OMNI-INTERACTSCOPE: BENCHMARKING PROACTIVE AND AGENTIC INTERACTION IN CONTINUOUS OMNIMODAL STREAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omnimodal large language models (Omni-LLMs) are increasingly expected to assist users while observing continuous audiovisual streams. However, existing evaluations provide limited insight into how reliably these models connect stream evidence and user needs to timely, sustained interventions. We introduce **Omni-InteractScope**, a benchmark of 1,500 human-annotated, cross-reviewed interactions with 1,814 temporal targets across eight task groups. It links input modalities, event triggers, intervention types, and user needs, covering event-timed tool invocation and query-free assistance alongside notifications, warnings, state tracking, and continuous responses. Our experimental results on the proposed benchmark reveal three critical limitations of current Omni-LLMs: (1) **Timely tool invocation remains unreliable**: all 12 evaluated models produce an on-time first response for fewer than 40% of tool-invocation targets, and joint timing-and-tool accuracy reaches at most 31.1%. (2) **Query-free reminders remain poorly covered**: all evaluated models achieve lower timely reference-target recall on query-free reminders than on explicit questions, measured on their respective target sets. Reminder recall never exceeds 23.0%. (3) **End-to-end guidance coverage remains limited**: nine models cover none of the 26 guidance sequences in full, and the remaining three cover only one to three sequences each. Targeted controls further show that tool-invocation scores alone do not establish event grounding and that reminder coverage depends on response history and frequency. These findings motivate developing and evaluating streaming assistants for timely action, initiative under general user needs, and coverage of entire response sequences.

1 INTRODUCTION

Omnimodal large language models (Omni-LLMs) such as GPT-4o, Qwen2.5-Omni, and Qwen3-Omni bring joint audiovisual understanding into conversational systems (Hurst et al., 2024; Xu et al., 2025a;b). Emerging streaming systems extend this ability to deciding when to speak, remain silent, or invoke tools (Chen et al., 2024; Wang et al., 2026b; ByteDance Seed, 2026). For an assistant observing a changing environment, these decisions must serve a user's ongoing needs: retrieve information when a requested event occurs, relay a relevant sound to a user with hearing impairment, or provide continue guidance as an activity unfolds (see Figure 1). Useful assistance therefore requires connecting perceptual evidence to *what intervention is needed, when it is useful, and whether further responses remain necessary*.

Evaluation has progressed from video comprehension toward interactive assistance. Early works, such as Video-MME and LongVideoBench, assess understanding of supplied videos (Fu et al., 2025; Wu et al., 2024), while StreamingBench and OVO-Bench restrict observations to available stream prefixes and test responses to unfolding events (Lin et al., 2026; Niu et al., 2025). Recent benchmarks further evaluate multimodal turn-taking, audio-dependent proactivity, and event-triggered response windows (Wang et al., 2025b; Zhao et al., 2026; Lu et al., 2026b). OmniAssistBench organizes goal-oriented assistance along predefined interaction paths (Sun et al., 2026). StreamArena jointly evaluates perception, retrospection, proactivity, and tool utilization over hour-scale causal streams (Zhang et al., 2026). Assistance without an event-specific question is also studied in VI-

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.

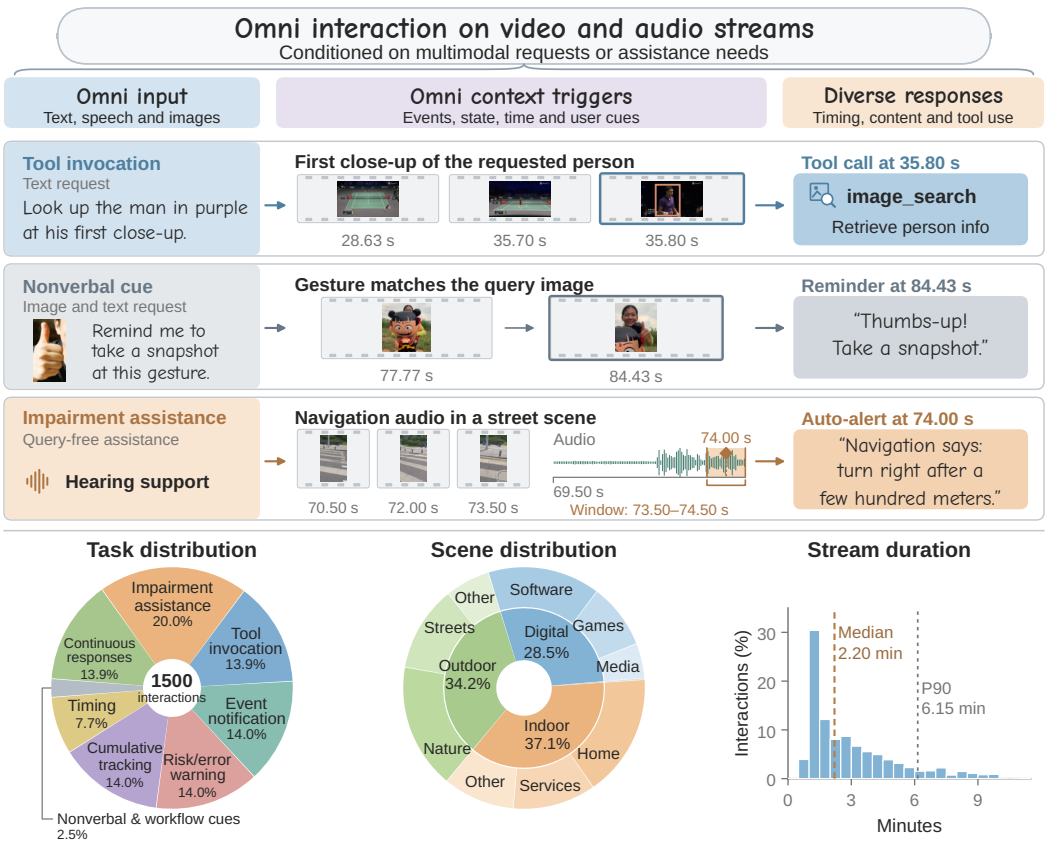


Figure 1: **Overview of Omni-InteractScope. Top:** Representative interactions connect user requests or general assistance needs with events in audiovisual streams and the content and timing of the required responses. The examples span event-timed tool invocation, a gesture reminder conditioned on an image and text request, and hearing assistance under a general instruction without an event-specific query. **Bottom:** Distributions of task types, scene categories, and stream durations across the benchmark's 1,500 interactions. Clips reused across multiple interactions are counted once for each interaction in which they appear.

ABench (Liu et al., 2026). Table 1 summarizes this progression in terms of stream access and intervention objectives. Section 2 provides a comprehensive discussion of related work.

An assistant's obligation extends beyond the content of its next answer. A specified event can require a tool call at a useful moment, a general assistance need can require intervention without an event-specific question, and a standing request can require further responses after an initial answer. Evaluating these requirements together calls for explicit links between user context, response content, valid timing, and successive targets. It also calls for distinguishing scene evidence from request priors and target coverage from the opportunities allowed to respond. This combination makes it possible to examine both whether an intervention obligation is fulfilled and which evaluation conditions shape the result.

We introduce **Omni-InteractScope**, a benchmark organized around the relationship between stream evidence, user needs, and required interventions. It contains 1,500 human-annotated, cross-reviewed interactions with 1,814 temporal targets. The assistance panel draws on 208 self-recorded first-person videos. The benchmark links four dimensions: input modalities, event triggers, intervention types, and user needs. Eight task groups span tool invocation, notifications, warnings, cumulative-state tracking, temporal conditions, nonverbal cues, continuous responses, and impairment assistance. Each target specifies response content and a valid time or interval. Interactions with multiple targets record the responses required after an initial answer. This organization supports evaluating both individual interventions and coverage of continuing obligations within a shared resource.

Within this broad coverage, we focus on two complementary demands of proactive interaction. Tool invocation requires selecting the appropriate tool and acting when a user-specified event occurs. Query-free assistance requires recognizing when a general user need warrants a reminder without an event-specific query. Paired controls examine the roles of scene evidence and response policy, while successive targets assess whether the assistant delivers the further responses required as an interaction unfolds.

Evaluation across 12 models reveals substantial limitations in both timely intervention and sustained assistance. Every evaluated model produces an on-time first response for fewer than 40% of tool-invocation targets, and joint accuracy reaches at most 31.1%. Timely reference-target recall on query-free reminders never exceeds 23.0%. Although guidance-step recall reaches 51.7%, nine models cover none of the 26 guidance sequences in full, and no model covers more than three.

Our main contributions are summarized as follows:

- We introduce Omni-InteractScope, a comprehensive benchmark for timely and sustained omnimodal intervention. It comprises 1,500 human-annotated, cross-reviewed interactions across eight task groups, with 1,814 targets specifying what to produce and when, including successive responses under continuing user needs.
- We provide focused evaluation of event-timed action and query-free initiative through 209 tool-invocation targets and 122 query-free reminders. Our diagnostics distinguish tool choice from timing and test whether invocation scores depend on scene evidence. Paired controls further examine how explicit event cues, response scheduling, history, and response frequency affect coverage on the same reminder targets.
- We perform a comprehensive evaluation of 12 multimodal models, providing empirical insights into timing, query-free assistance, and sequence coverage, exposing gaps between tool selection and timely action, question answering, and query-free reminders, and step coverage and complete guidance.

2 Related Work

From video understanding to proactive interaction. Video-MME and LongVideoBench evaluate comprehension of supplied videos (Fu et al., 2025; Wu et al., 2024). StreamingBench and OVO-Bench restrict observations to available stream prefixes and test responses to unfolding events (Lin et al., 2026; Niu et al., 2025). OmniMMI and OmniPro extend evaluation to multimodal interaction and audio-dependent proactivity (Wang et al., 2025b; Zhao et al., 2026). IPIBench studies evolving requests, while OmniInteract evaluates turn-taking, interruption, and recovery (Li et al., 2026; Lu et al., 2026b). These works motivate causal evaluation of response content and timing. Omni-InteractScope records the intervention required by a user need, including targets that remain after an initial response.

Tool use and continuing response obligations. ToolBench and API-Bank evaluate tool-use plans and executable calls (Qin et al., 2024; Li et al., 2023). StreamArena brings tool utilization and proactivity into hour-scale causal streams, including external search and answer integration (Zhang et al., 2026). Our T1 setting evaluates tool identity and invocation timing, with controls that distinguish request priors from scene evidence. Omni-DuplexEval includes continuous descriptions and event-conditioned reminders, and OmniAssistBench evaluates assistance along predefined paths (He et al., 2026; Sun et al., 2026). We retain successive targets within each interaction and report both individual-target and complete-guidance coverage. Table 1 separates these objectives from multiple independent questions about the same video.

Assistance under general user needs. VizWiz studies visual questions from blind users (Gurari et al., 2018), while VIABench evaluates reminders, questions, and guided interaction (Liu et al., 2026). Vinci2 studies assistance without explicit prompts, and SoundWatch relays environmental sounds to deaf and hard-of-hearing users (Gong et al., 2026; Jain et al., 2020). Our assistance panel includes visual- and hearing-assistance needs, distinguishing query-free reminders from explicit questions and multi-step guidance. Following concerns about response frequency in SPOT-Bench (Chatterjee et al., 2026), paired controls examine how history and generation opportunities affect reminder coverage and output volume on the same episodes.

Table 1: **From video understanding to streaming intervention.** Coverage of scene inputs and interaction objectives at benchmark level.

Resource			Evaluation		Intervention objectives			
Benchmark	Modality	#Videos	Open-ended	Streaming	Proactive	Tool	Query-free	Successive
(A) Offline video understanding								
Video-MME (Fu et al., 2025)	V, A, S	900	–	–	–	–	–	–
LongVideoBench (Wu et al., 2024)	V, S	3,763	–	–	–	–	–	–
(B) Streaming understanding and interaction								
StreamingBench (Lin et al., 2026)	V, A	900	–	✓	✓[p]	–	–	–
OVO-Bench (Niu et al., 2025)	V	644	Mix	✓	✓[p]	–	–	✓
OmniMMI (Wang et al., 2025b)	V, A	1,121	Mix	✓	✓	–	–	–
OmniPro (Zhao et al., 2026)	V, A	1,262	✓	✓	✓	–	–	✓
IPIBench (Li et al., 2026)	V	1,831	✓	✓	✓	–	–	✓
OmniInteract (Lu et al., 2026b)	V, A	250	✓	✓	✓	–	–	✓
OmniAssistBench (Sun et al., 2026)	V, A	300	✓	✓	✓	–	–	✓
Omni-DuplexEval (He et al., 2026)	V, A	660	✓	✓	✓	–	–	✓
StreamArena (Zhang et al., 2026)	V, A	243	✓	✓	✓	Exec.	–	–
(C) Assistance without an event-specific user query								
VIABench (Liu et al., 2026)	V	761	✓	✓	✓	–	✓	✓
Omni-InteractScope (ours)	**V, A**	**1,128[r]**	✓	✓	✓	Call	✓	✓

V/A/S: video/audio/subtitles; queries are omitted. Audiovisual input does not establish audio necessity. –: not an explicit objective; Mix: includes multiple-choice tasks. Streaming includes causal prefix/replay, not necessarily native duplex inference; [p]proactivity uses repeated external probes. Call: tool identity and timing; Exec.: execution and answer integration. Query-free: no event-specific query. Successive: multiple targets under one continuing need. [r]Video records, including 208 self-recorded assistance videos.

3 OMNI-INTERACTSCOPE

Omni-InteractScope comprises 1,500 interactions across eight task groups, organized around the interventions an assistant should deliver as a stream unfolds. Figure 1 summarizes the design, and Figure 2 illustrates each task group in details.

3.1 DESIGNING FOR TIMELY AND SUSTAINED INTERVENTION

The benchmark unit is an interaction: a video and audio stream, user context, and one or more reference interventions. Each intervention is a target specifying the required response and its reference time or valid interval. For a continuing request, the interaction retains successive targets in temporal order.

We organize this unit along four linked dimensions. **Input modalities** describe the available video, scene audio, and user context conveyed through text, speech, or image attachments. **Event triggers** describe the evidence or condition associated with a response, such as an object appearing, a sound, an accumulated count, or elapsed time. **Intervention types** specify what the assistant should produce: an answer, notification, warning, tool call, or successive commentary or guidance. **User needs** determine why that intervention matters, through an explicit request or a general visual- or hearing-assistance instruction.

Together, these dimensions connect stream evidence and user context to the interventions an assistant should provide. The eight task groups organize these requirements into a comprehensive evaluation of response content, timing, and coverage, encompassing both individual interventions and successive responses under continuing user needs.

3.2 TASK COVERAGE AND FOCAL SETTINGS

Responding to explicit requests. Seven task groups contain 1,200 interactions with different monitoring demands. **Event notification (T2)** asks the assistant to report a specified occurrence, while **risk/error interjection (T3)** calls for a warning when an unsafe or incorrect action occurs. **Nonverbal/workflow cues (T6)** connect a gesture or workflow event to the requested response,

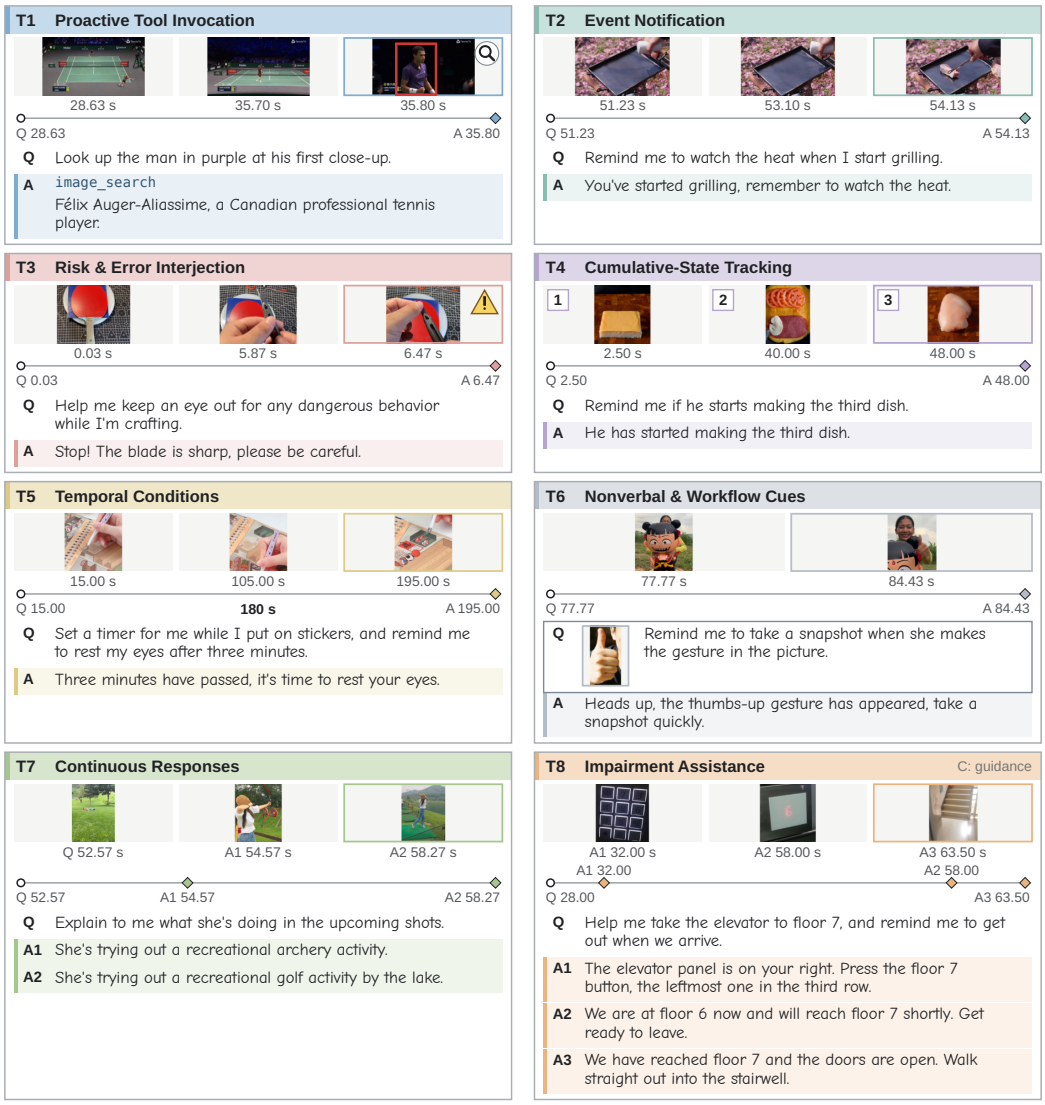


Figure 2: **Representative interactions across the eight task groups of Omni-InteractScope.** Each panel links a user request (Q), timestamped stream observations, and reference responses (A) to illustrate what intervention is required and when. The colored diamonds mark reference response times. T6 includes an image and text request, while T7 and the T8-C guidance example show how a single request can require successive responses as the stream unfolds.

including matching a gesture to a query image. **Cumulative-state tracking (T4)** uses earlier observations to interpret current events. **Temporal conditions (T5)** introduce obligations governed by elapsed time. **Continuous responses (T7)** extend a standing request across successive commentary or translation segments.

Event-timed tool invocation. T1 tests whether a model invokes the appropriate tool when a specified event occurs. A request can indicate which tool to use before the stream establishes when to act. Each target therefore specifies a tool-name set and invocation time. Evaluation covers these requirements, with arguments and downstream execution outside its scope.

Assistance under explicit and general needs. The 300-interaction **impairment-assistance panel (T8)** covers visual- and hearing-assistance needs. **Query-free reminders (T8-A)** require identifying useful interventions under a general instruction, without an event-specific question. **Audiovisual**

questions (T8-B) test answering explicit requests, while **multi-step guidance (T8-C)** tests successive responses along recorded trajectories. Both T7 and T8-C retain ordered targets within each interaction to represent continuing response obligations. Appendix A.2 details the category boundaries.

3.3 Data construction and annotation

The benchmark combines 920 online video records for explicit-request tasks with 208 self-recorded first-person video records for impairment assistance. It spans indoor, outdoor, and digital-screen settings, with QA text balanced between English and Chinese (750 interactions each). The 1,500 interactions contain 1,814 targets, with multiple targets in 165 interactions. Appendix A details task counts, scene and trigger distributions, duration statistics, and source provenance, including repeated source identities among video records.

Human annotators specified user context, reference responses, triggers, and response times or valid intervals. A separate group of human reviewers cross-checked the annotations to obtain the final benchmark. This review covers both response content and its temporal relation to the event. Appendix A.4 documents the supplemental consistency audit and review records.

4 Experiments

4.1 Experimental setup

Models. We evaluate 12 model configurations in three groups. (1) *Proprietary models* include Gemini-3.1-Pro (Google DeepMind, 2026a) and Gemini-3.5-Flash (Google DeepMind, 2026b), and Qwen3.5-Omni-Plus (Team, 2026). (2) *Open-source prompted MLLMs* include Qwen3-Omni (Xu et al., 2025b), Qwen2.5-Omni (Xu et al., 2025a), Qwen3-VL (Bai et al., 2025), and InternVL3.5 (Wang et al., 2025a). (3) *Streaming models* include LiveCC (Chen et al., 2025), MiniCPM-o 4.5 (Cui et al., 2026), MMDuet2 (Wang et al., 2026b), MOSS-VL (Wang et al., 2026a), and AURA (Lu et al., 2026a).

Response collection. Models receive causal inputs at query time, subsequent integer-second media times, and clip end; T8-A begins at stream start. The schedule is target-independent. *Offline Probing* uses independent prefixes without prior responses. *Online Rolling* replays pre-query observations with the request supplied, then retains native state across successive observations. Initialization outputs are excluded. Appendices B and C detail collection and adapters.

Scoring. Interaction i contains K_i targets, each specifying required content y_{ik} and a valid time window W_{ik}. For response j with content $\hat{y}_{ij}$ and timestamp t_{ij}, joint content–time eligibility is

$$E_{ikj} = \mathbb{I}[t_{ij} \in W_{ik}]\, C_i(\hat{y}_{ij}, y_{ik}), \tag{1}$$

where $\mathbb{I}$ is the indicator function and C_i returns 1 for correct content given the user context and 0 otherwise. T1 requires an exact match of tool-name sets, with no missing or extra tools. T2–T8 use a binary GPT-4o judge (Hurst et al., 2024) comparing the response with the request and reference, accepting equivalent wording. Timing is checked separately by the scorer.

T1–T7 and T8-B/C use a two-second tolerance around the reference time, with T8-B/C lower bounds clipped to query time. T8-A uses its annotated valid interval. All timestamps use media time, excluding inference latency. Appendix B specifies timestamp conventions and window boundaries.

T1–T7 score the first response assigned to each target. A premature or incorrect response cannot be rescued by a later correction. T8 instead maximizes content–time matches while preserving target and response order, using each response at most once. Unmatched targets and incomplete interactions score zero.

Metrics. Let $m_{ik} \in \{0, 1\}$ indicate target success under these rules. For a reported set of interactions $\mathcal{S}$, we aggregate target successes as

$$S(\mathcal{S}) = \frac{\sum_{i \in \mathcal{S}} \sum_{k=1}^{K_i} m_{ik}}{\sum_{i \in \mathcal{S}} K_i}. \tag{2}$$

Table 2: **Main results on explicit-request tasks (T1–T7).** Joint content–timing accuracy (%). Avg. equally weights T1–T4, T6, and T7.

Model configuration			Joint content–timing accuracy (%) ↑							
Model	**Size**	**Input**	**T1** Tool	**T2** Event	**T3** Risk	**T4** State	**T5*** Time	**T6** Cue	**T7** Cont.	**Avg.** 6 tasks
Targets			209	209	209	210	115	37	462	
Proprietary Models (prompted)										
Gemini-3.1-Pro	N/R	V, A	22.0	27.8	12.0	**11.0**	N/A	**27.0**	11.0	18.5
Gemini-3.5-Flash	N/R	V, A	28.7	**31.1**	**14.4**	10.0	N/A	24.3	**11.3**	**20.0**
Qwen3.5-Omni-Plus	N/R	V, A	**31.1**	23.9	4.8	3.8	N/A	10.8	7.8	13.7
Open-source MLLMs (prompted)										
InternVL3.5	8B	V+ASR	28.2	1.0	0.0	0.0	N/A	5.4	0.4	5.8
Qwen3-VL	8B	V+ASR	19.1	4.8	0.5	0.0	N/A	2.7	3.0	5.0
Qwen2.5-Omni	7B	V, A	10.5	0.5	0.5	0.5	N/A	0.0	1.5	2.2
Qwen3-Omni	30B-A3B	V, A	25.8	2.9	1.4	1.0	N/A	2.7	5.6	6.6
Streaming Models (rolling)										
LiveCC	7B	V	0.0	1.4	0.5	0.0	0.0	0.0	0.6	0.4
AURA	8B	V	0.0	4.8	0.5	1.4	0.0	2.7	0.6	1.7
MMDuet2	3B	V	0.0	3.8	1.0	0.5	0.0	2.7	1.7	1.6
MOSS-VL	11.3B	V	1.4	9.6	1.9	5.7	0.9	8.1	5.8	5.4
MiniCPM-o 4.5	9B	V, A	1.4	8.1	1.4	1.0	1.7	2.7	1.1	2.6

Notes. Prompted models use *Offline Probing* with independent causal-prefix calls, while streaming models use *Online Rolling* with retained state (Section 4.1). *T5 retains the earlier query-relative rolling evaluation as a separate diagnostic, excluded from every Avg. N/A denotes probing inputs without an explicit elapsed-time clock. V: video, A: native audio, ASR: scene transcripts. A3B denotes active parameters, and N/R represents undisclosed size. Bold marks column maxima in the six-task comparison.

We report S as task accuracy for T1–T7 and timely reference-target recall for T8, in percentages. Unmatched T8 outputs incur no penalty, so recall does not measure selectivity or interruption cost. Complete T8-C counts interactions with every guidance step covered, out of 26 sequences. Appendix B gives formal metrics, content rubrics, and assignment rules.

4.2 Main results

Tables 2 and 3 report explicit-request and assistance results. Avg. equally weights T1–T4, T6, and T7, with T5 reported separately. Assistance results pool target recall and count complete guidance sequences. Configurations differ in inputs and execution interfaces (Appendix D.4.1 examines within-model schedule effects). After media-validity exclusions, evaluation covers 1,498 interactions and 1,812 targets (Appendix A.1).

Task strengths vary across intervention types. Gemini-3.5-Flash leads the six-task Avg., while Qwen3.5-Omni-Plus leads tool invocation. Yet the best scores on cumulative-state tracking (T4) and continuous responses (T7) are only 11.0% and 11.3%, respectively. These weaknesses span both timing and content: Gemini-3.5-Flash responds prematurely on 176/210 T4 targets, while its T7 failures include 234/462 premature responses and 156/462 timely content failures (Appendix D.4.3). Task-level and failure breakdowns therefore reveal limitations obscured by the aggregate ranking.

Reminder coverage is lower than question coverage. Every configuration covers a smaller fraction of query-free reminders (T8-A) than explicit questions (T8-B). Even Gemini-3.1-Pro, which leads both, reaches 23.0% on reminders versus 62.5% on questions. These are different event sets. Section 5.2 tests guidance and response policies on matched reminder episodes.

Step coverage does not ensure complete guidance. Gemini-3.1-Pro covers 51.7% of T8-C steps but completes only 3 of 26 guidance sequences. No open-source prompted or streaming configuration completes a sequence. Complete T8-C measures coverage of the annotated trajectory, not closed-loop user-task success. Section 5.3 relates completion to individual-step coverage.

Table 3: **Main results on assistance tasks (T8).** Timely reference-target recall (%) for query-free reminders (T8-A), audiovisual QA (T8-B), and guidance steps (T8-C), with 122/152/87 targets (361 overall). Overall pools hits across all targets. Complete T8-C counts fully covered guidance sequences out of 26. All configurations use full-clip causal collection, response-availability timestamps, and ordered matching (Section 4.1). Bold marks column maxima.

Model configuration			Timely target recall (%) ↑				Coverage
Model	**Size**	**Input**	**Overall**	**T8-A**	**T8-B**	**T8-C**	**Complete T8-C**
Proprietary Models							
Gemini-3.1-Pro	N/R	V, A	**46.54**	**22.95**	**62.50**	**51.72**	**3/26**
Gemini-3.5-Flash	N/R	V, A	41.27	22.13	57.89	39.08	2/26
Qwen3.5-Omni-Plus	N/R	V, A	31.30	9.02	50.00	29.89	1/26
Open-source MLLMs							
InternVL3.5	8B	V+ASR	14.40	0.00	26.97	12.64	0/26
Qwen3-VL	8B	V+ASR	14.96	2.46	24.34	16.09	0/26
Qwen2.5-Omni	7B	V, A	10.25	2.46	20.39	3.45	0/26
Qwen3-Omni	30B-A3B	V, A	13.02	2.46	25.66	5.75	0/26
Streaming Models							
LiveCC	7B	V	3.60	0.00	8.55	0.00	0/26
AURA	8B	V	6.93	0.00	15.79	1.15	0/26
MMDuet2	3B	V	1.94	0.00	4.61	0.00	0/26
MOSS-VL	11.3B	V	3.32	0.00	7.24	1.15	0/26
MiniCPM-o 4.5	9B	V, A	3.05	0.00	7.24	0.00	0/26

5 Diagnostic Analysis

5.1 Scene evidence and tool-invocation scores

Figure 3a shows that removing scene input does not lower Gemini-3.5-Flash's T1 accuracy. On the same 209 interactions, joint accuracy is 28.71% with scene input and 32.06% without it, with requests and the evaluation protocol held fixed. The paired difference is -3.35 percentage points. Joint accuracy alone therefore does not establish event-grounded invocation. Appendices D.1.2 and D.4.2 detail the paired ablation and temporal-tolerance analysis.

The request-only control reveals strong task priors. Gemini-3.5-Flash selects 96.7% of tools from request text and public definitions alone. Emitting these choices after a fixed three-second delay gives 67.5% joint success, the highest of four tested delays. This separate control uses a closed-label output interface and different response opportunities from the main evaluation, so it diagnoses request and timing priors rather than providing a matched model ranking (Appendix D.1.1).

5.2 Reminder coverage under different response policies

As shown in Figure 3b, giving Gemini-3.5-Flash response history reduces output volume by 58.1%, while both conditions cover 22 of 122 T8-A targets. The matched targets change, with 12 gained and 12 lost. Adding a cooldown reduces outputs further but costs 9.84 percentage points of recall relative to history alone. The policy is implemented by skipping generation calls during the cooldown, restricting the opportunities to cover reminder targets (Appendix D.2.2).

Figure 3c compares event-specific and general instructions in separate runs on the same 122 T8-A episodes. Gemini-3.5-Flash and Qwen2.5-Omni benefit from event cues when a response is requested at an annotated valid time. Under autonomous responding, however, the effects are small and uncertain. The benefit of event cues is therefore clearer when the response time is supplied than when the model must decide when to intervene. Scheduling changes temporal information and response opportunities together, so this comparison does not isolate a planning mechanism.

5.3 Complete guidance relative to step coverage

On the same 20 three-step T8-C episodes, Gemini-3.1-Pro covers 30 of 60 steps but completes only three sequences. Complete T8-C complements step recall by recording whether all required guidance was delivered.

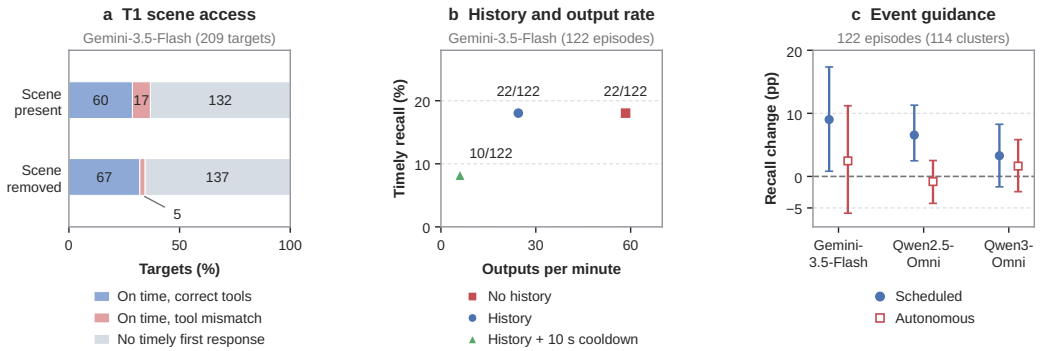


Figure 3: **Scene access and response opportunities shape intervention scores.** (a) T1 outcomes for Gemini-3.5-Flash with and without scene input (209 targets, two-second tolerance). Labels report target counts. (b) T8-A timely recall versus output rate for Gemini-3.5-Flash with no history, the five most recent responses, or history plus a ten-second cooldown. (c) Change in T8-A recall from general to event-specific instructions under scheduled and autonomous responding. Error bars show 95% paired cluster-bootstrap confidence intervals. Appendices D.1.2, D.2.2, and D.2.1 detail the protocols and results.

Figure 6 (Appendix D.3) compares observed completions with a reference based on success rates at each step position. An average of 2.4 complete sequences is expected if successes at different step positions are independent (95% reference interval: $[0, 5]$). The observed count is consistent with this reference and does not establish a separate memory or planning failure.

6 Discussion

Invocation scores need evidence controls. The T1 controls show that request information and response timing can sustain joint success without scene inputs. A tool-and-time score should therefore be accompanied by evidence controls when used to assess event grounding. For streaming tool use (Zhang et al., 2026), varying event occurrence and timing while holding requests fixed would test a stronger requirement than the scene ablation evaluated here.

Coverage depends on the response policy. Response history reduces Gemini-3.5-Flash's output volume with the same aggregate recall point estimate, while a cooldown reduces both outputs and coverage. These findings connect response-frequency evaluation (Chatterjee et al., 2026) to the policies used during generation. Reporting history access and output rate alongside recall makes this dependence visible. Selectivity additionally requires evaluating the value and cost of interventions beyond the annotated targets, an important direction for future benchmark development.

Complete coverage has a separate measurement role. Users may need every guidance step even when individual responses are often correct. Complete T8-C expresses that requirement, while the sequence reference limits attribution of low completion to a separate memory or planning failure. For goal-oriented assistants (Sun et al., 2026), explicit records of delivered guidance and pending obligations remain a design hypothesis to test through interventions on history and state.

7 Conclusion

We introduced Omni-InteractScope, a benchmark for timely and sustained intervention in continuous omnimodal streams. Its 1,500 human-annotated interactions span eight task groups, linking user needs and stream evidence to 1,814 targets that specify response content and timing. Evaluation across 12 models reveals substantial limitations in timely tool invocation, query-free reminders, and complete guidance coverage. These findings underscore the need to assess whether assistants fulfill the response requirements of an interaction, including interventions that remain necessary after an initial answer. Omni-InteractScope provides a shared basis for developing and evaluating assistants that respond to unfolding events and sustain support for continuing user needs.

AI USE STATEMENT

In this work, we used generative AI tools to help reorganize the benchmark task taxonomy, support translation, and improve the clarity and language of the manuscript. We also used an LLM judge for preliminary data validation prior to final human verification. Separately, LLM-based content judging was used in benchmark evaluation, as detailed in Appendix B. AI tools additionally assisted with organizing and presenting qualitative cases in Appendix E. Generative AI was not used to generate synthetic benchmark data, propose or refine research hypotheses, design the research methodology or experiments, implement the evaluated methods, or clean or reformat the benchmark dataset. It was also not used to develop theoretical models or new conceptual frameworks, formulate mathematical claims, or contribute to or write mathematical proofs. The authors retained responsibility for methodological decisions, qualitative judgments, and the scientific interpretation of the results. All benchmark media and annotations underwent rigorous human review. We take full responsibility for the final content of this work, including any text, claims, code, or other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

Sections 3 and 4.1 describe benchmark design, data construction, the experimental setup, and the primary evaluation metrics. Appendix A details dataset composition, taxonomy boundaries, annotation procedures, and consistency checks. Appendix B specifies observation schedules, timestamp conventions, temporal eligibility, response assignment, and uncertainty estimation, with worked examples of both matching rules. Appendix C provides input adapters, decoding configurations, and core prediction and judging prompts. Appendix D documents control conditions, sample sizes, and sensitivity analyses supporting the diagnostic findings. Appendix E documents source frames, recorded model outputs, timing information, and scoring context for five illustrative cases.

REFERENCES

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.

ByteDance Seed. SeedRealtime Audio-Visual Full-Duplex LLM Released: Toward Omni-Modal Natural Interaction, Aug 2026. URL `https://seed.bytedance.com/en/blog/seedrealtime-audio-visual-full-duplex-llm-released-toward-omni-modal-natural-interaction`. ByteDance Blog.

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.

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.

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.

Sitong Gong, Tianyu Yan, Caixin Kang, Bo Zheng, Xiang Ruan, Huchuan Lu, Kaipeng Zhang, Yoichi Sato, and Yifei Huang. Vinci2: Providing proactive assistance in continuous egocentric videos. *arXiv preprint arXiv:2607.11523*, 2026.

Google DeepMind. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026a. Accessed: 2026-08-21.

Google DeepMind. Gemini 3.5 flash model card. https://deepmind.google/models/model-cards/gemini-3-5-flash/, 2026b. Accessed: 2026-08-21.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617. IEEE, 2018.

Chaoqun He, Mingyang Xiang, Yingjing Xu, Bokai Xu, Junbo Cui, Jie Zhou, Yuan Yao, and Lijie Wen. Omni-duplexeval: Evaluating real-time duplex omni-modal interaction. *arXiv preprint arXiv:2605.17360*, 2026.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Dhruv Jain, Hung Ngo, Pratyush Patel, Steven Goodman, Leah Findlater, and Jon Froehlich. Soundwatch: Exploring smartwatch-based deep learning approaches to support sound awareness for deaf and hard of hearing users. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 1–13, 2020.

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.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. Api-bank: A comprehensive benchmark for tool-augmented llms. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 3102–3116, 2023.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Yunfeng Liu, Yuandong Yang, Jiarui Han, Zhenpeng Huang, Yuqing Tang, Xiangyu Zeng, Gangshan Wu, and Limin Wang. Viabench: A comprehensive video benchmark collected from blind individuals for visual impairment assistance. *arXiv preprint arXiv:2607.14660*, 2026.

Xudong Lu, Yang Bo, Jinpeng Chen, Shuhan Li, Xintong Guo, Huankang Guan, Fang Liu, Dunyuan Xu, Peiwen Sun, Heyang Sun, et al. Aura: Always-on understanding and real-time assistance via video streams. *arXiv preprint arXiv:2604.04184*, 2026a.

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*, 2026b.

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.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

Xianyun Sun, Chaoyou Fu, Zhengye Zhang, Feiyang Duan, Qingyuan Cao, Yonghui Niu, Sihang Yuan, Ge Zhang, and Caifeng Shan. Omniassistbench: Assistant-style interaction benchmark for omni-llms. *arXiv preprint arXiv:2608.21360*, 2026.

Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Pengyu Wang, Chenkun Tan, Shaojun Zhou, Qirui Zhou, Yanxin Chen, Xingyang He, Huazheng Zeng, Jijun Cheng, Chenghao Wang, Xiaomeng Qian, et al. Moss-vl technical report. *arXiv preprint arXiv:2608.15045*, 2026a.

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

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, 2026b.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omnimmi: 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.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

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. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

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. *arXiv preprint arXiv:2608.05703*, 2026.

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*, 2026.

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

APPENDIX GUIDE

Appendix A documents data, taxonomy, and annotation. Appendix B specifies observation and scoring rules, including worked examples. Appendix C gives model settings, adapters, and core prompts. Appendix D expands the three experimental questions and tests evaluation sensitivity. Appendix E examines actual model responses.

A DATASET, TAXONOMY AND ANNOTATION

A.1 INTERACTION RECORDS AND TASK COVERAGE

The coverage statistics below expand the interaction and target representation in Section 3, preserving successive targets within each interaction.

Table 4: Task coverage and annotated trigger modalities. V, A, and AV denote visual, auditory, and joint audiovisual triggers. Txt and Oth denote textual and other triggers. The five columns partition the reference targets within each task.

Task	Behavior	Interactions	Targets	V	A	AV	Txt	Oth
T1	Tool invocation	209	209	198	10	1	0	0
T2	Event notification	210	210	153	36	18	3	0
T3	Risk/error interjection	210	210	208	0	2	0	0
T4	Cumulative-state tracking	210	210	175	3	6	3	23
T5	Temporal conditions	115	115	36	0	0	0	79
T6	Nonverbal/workflow cues	37	37	32	0	4	1	0
T7	Continuous responses	209	462	379	62	16	5	0
T8-A	Query-free reminders	122	122	101	16	5	0	0
T8-B	Audiovisual QA	152	152	136	11	5	0	0
T8-C	Guided interaction	26	87	79	7	1	0	0
Total		**1,500**	**1,814**	**1497**	**145**	**58**	**12**	**102**

Table 4 expands Figure 1. T1–T7 contain 1,200 interactions and 1,453 targets, and T8 contains 300 interactions and 361 targets. Multiple targets occur in 165 interactions. T1 has one target per interaction, whereas T7 has 462 targets in 209 interactions and T8-C has 87 in 26.

Evaluation population. Two single-target interactions, one each in T2 and T3, fail the media-validity check. The resource retains 1,500 interactions and 1,814 targets, while evaluation uses 1,498 interactions and 1,812 targets: 1,083 interactions and 1,336 targets in T1–T4, T6, and T7, 115 in the separate T5 diagnostic, and all 300 T8 interactions with 361 targets. Incomplete model trajectories retain their targets in these denominators (Appendix B.3).

Request formats are text (1,223 interactions), audio (124), text with an image (25), text with audio (6), and no event-specific query (122, exactly T8-A). Request modality and trigger modality describe different parts of the interaction: a visual event can, for example, indicate when to search the current audio. Actual model inputs are listed in Table 6.

A.2 TAXONOMY DEFINITIONS AND BOUNDARIES

Trigger labels describe annotated evidence rather than establishing that a modality is indispensable under every input condition. Intervention type describes the output, while user need explains its purpose. Similar outputs can therefore arise from explicit event requests or general assistance instructions, with different requirements for initiative.

Requested actions and notifications (T1–T3). T1 requires invoking the reference tool set when the event occurs; tool identity and timing are scored, while arguments and execution are outside the

metric. T2 reports a requested occurrence. T3 additionally requires communicating an unsafe or incorrect action: merely announcing the event does not convey the required warning.

What makes a response due (T4–T6). T4 depends on accumulated observations, such as the start of a third dish after two earlier preparations. T5 depends on an elapsed duration and is reported separately. T6 responds to a gesture or workflow cue, which can require matching the observed gesture to a query image. A query attachment specifies the sought cue; the stream supplies evidence that it has occurred.

General assistance and explicit questions (T8-A/B). T8-A supplies a general visual- or hearing-assistance need without an event-specific question. The model must identify a useful reminder from the stream. T8-B instead names the immediate information need through an explicit audiovisual request. Similar response wording can therefore reflect different requirements for initiative.

Continuing responses and actionable guidance (T7, T8-C). Both contain successive targets under one standing need. T7 covers continuing commentary or translation. T8-C requires actionable instructions justified by successive states along a recorded assistance trajectory and defines the unit for complete-guidance reporting. T1–T7 group interventions and triggering conditions, whereas T8 groups assistance needs; timely action and sustained coverage can consequently apply to several task groups.

A.3 Sources and annotation

The explicit-request panel contains 920 video records: 903 from Douyin, 11 from Xiaohongshu, and six from Bilibili, with 503 first-person and 417 third-person records. The assistance panel uses 208 self-recorded first-person videos for 262 visual-impairment and 38 hearing-impairment interactions. QA text is balanced between English and Chinese at 750 interactions each; scene audio retains its original language. Indoor, outdoor, and digital-screen settings account for 37.1%, 34.2%, and 28.5% of interactions. Median clip duration is 2.20 minutes, and the 90th percentile is 6.15 minutes. Figure 1 counts each interaction in these distributions and uses 30-second duration bins.

The 1,128 video records are storage and analysis units, not necessarily distinct original sources. Seven groups in the explicit-request panel share platform identities. Uncertainty estimates use video-record clusters; the source inventory and identifier discrepancies are documented in the provenance materials.

Construction proceeds from source records to interaction annotations and then cross-review. For each interaction, annotators record the user request or general need, query time and attachments where applicable, and the required interventions. Each target retains its content, triggering evidence, and time or interval. Related targets remain ordered in one interaction, preserving the distinction between a continuing need and independent questions. A separate group of human reviewers cross-checks the annotations to produce the final benchmark.

A.4 Annotation consistency checks

A stratified annotation-metadata audit contains 64 primary reviews, split evenly between T1 and T8-A, and 16 secondary reviews. Sampling covers less frequent tools and trigger modalities without using evaluated-model predictions. Final records mark evidence support positive in 62/64 cases, timing in 63/64, modality in 62/64, and tool or assistance need in 63/64. Five cases include documented Gemini-3.5-Flash-assisted review with human confirmation.

A separate trigger-label check used Gemini-3.5-Flash and Gemini-3.1-Pro on six targets from four videos, providing both models with the same video crops, original audio, and annotation context. Both assigned all six triggers to the visual category. These labels are incorporated into Table 4 and the modality analysis; reference content and timing are unchanged. Individual judgments, adjudication notes, and revision records are retained with the audit.

A.5 From scene evidence to an annotation record

Figure 2 pairs source frames with requests and reference responses, including accumulated state, elapsed time, query-image cues, and successive targets. Figure 1 additionally illustrates query-free hearing assistance with original scene audio represented by its measured energy envelope. Figure 4 expands the cumulative-state example to show how the request, evidence, target time, and response are represented. The two-second tolerance is applied by the evaluator to a point annotation; it is not an additional human-annotated interval.

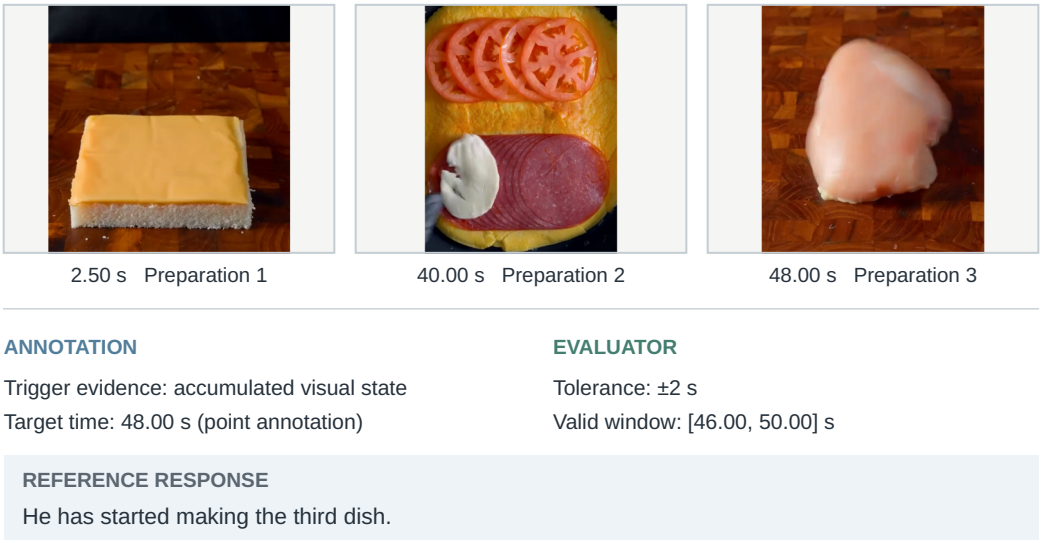


Figure 4: **An annotated cumulative-state interaction (T4).** Three source frames show successive dish preparations. The request and reference response are reproduced from the annotation. Count labels identify the illustrated preparations, not model predictions. The evaluator derives $[46, 50]$ seconds from the 48-second point target. The frame strip provides selected evidence; the model observes the causal stream rather than this summary.

B Evaluation Protocol and Scoring

B.1 Observation schedules and response history

Main decision grid. For T1–T4, T6, T7, and T8, decisions occur at $\{q, D\} \cup \{k \in \mathbb{Z} : q \leq k \leq D\}$, where q is query time and D is clip duration, restricted to available media. A noninteger query time is included before the next integer tick. T8-A starts at stream onset, and T8 decisions require an available video frame. Models receive no reference answers, target times, or annotation-derived end times. The scheduled T8 controls are the explicit exception: they request one response at an annotated valid time (Appendix D.2.1).

Prompted models make independent prefix calls without previous responses. Stateful adapters retain a session, initialized at query arrival by replaying earlier observations with the request already supplied. Initialization outputs are excluded from evaluation; native state, including generated history, is retained. Thus the main comparison shares decision opportunities and access to prior observations, while input representation and memory mechanisms differ. The additional Gemini-3.5-Flash history controls use their own regenerated baseline (Appendix D.2.2).

Execution may stop after the final target deadline only when earlier outputs are invariant to stopping. All earlier decision opportunities are retained. MiniCPM-o 4.5 and MOSS-VL run in full, and LiveCC must cover the entire last response group.

Separate T5 diagnostic. T5 retains 115 rolling interactions beginning at query time without historical initialization, with observations at $q, q+1, q+2, \ldots$ and adapter onset timestamps. These results use a different schedule from the main six-task comparison. Prompted T5 inputs lack an explicit elapsed-time clock, so their cells are N/A. Every model's Avg. excludes T5.

Response units and timestamps. All times refer to the media clock. For T1–T4, T6, and T7, prompted responses use prefix decision time. MiniCPM-o 4.5 and MOSS-VL use saved turn or sentence onset; MMDuet2 and AURA return within one tick, making onset and collection time equal. LiveCC chunks are assembled within two seconds of the first post-query chunk and credited at that chunk's time, without chaining consecutive gaps. Responses with saved onset before the query are excluded. For T8, responses use collection time; LiveCC continuations are joined at the utterance boundary or stream end and credited at the last chunk. This measures when assistance content becomes available, rather than wall-clock serving latency.

B.2 Notation and temporal eligibility

We record response j in interaction i as $(t_{ij}, \hat{y}_{ij})$, with a media timestamp and response content. Interaction i contains K_i reference targets. Each target k specifies required content y_{ik} and an acceptable interval W_{ik}. Let $C_i(\hat{y}, y_{ik}) \in \{0, 1\}$ denote content correctness given the user context. Content and time are checked separately.

T1–T7 accept timestamps in $W_{ik} = [\tau_{ik} - 2, \tau_{ik} + 2]$ seconds around the annotated target time τ_{ik}. T8-A uses the annotated valid interval without extension. T8-B and T8-C use $W_{ik} = [\max(q_i, \tau_{ik} - 2), \tau_{ik} + 2]$, where q_i is query time. All endpoints are inclusive. Equation 1 defines E_{ikj} as joint content–time eligibility, using the indicator function $\mathbb{I}[\cdot]$. Output timestamps follow the response units and media-time conventions in Appendix B.

B.3 Content criteria and judge inputs

T1 compares the predicted and reference tool-name sets, accepting supported formatting variations while requiring no missing or extra tools. It does not score tool arguments, execute tools, or judge retrieved answers.

For T2–T8, the content judge is GPT-4o, with temperature zero and a ten-token verdict budget. It receives the request, reference, and candidate response, without scene media or timestamps. T8 adds the target user group and subtask to the request context. Timing is checked separately by the scorer. Table 5 summarizes the criteria used by the recorded prompts, following rubric-based LLM evaluation (Zheng et al., 2023; Liu et al., 2023).

Table 5: Content criteria. Equivalent wording is accepted. Each rubric evaluates correspondence to the reference target, independently of response timing.

Rubric	Required content	Failure examples
General	Correct facts, names, quantities, and yes/no decisions.	Wrong facts, hallucinations, or refusals.
Event	The same event with its key subject, object, and action.	Generic narration, an unrelated event, or an incorrect participant or action.
Translation	The reference meaning, allowing synonyms and differences in fluency.	Omitted meaning or added content.
Assistance	The target information and any essential action. A guidance step must state the currently actionable instruction.	Contradictory directions, unsafe advice, invented critical details, a generic warning, or repetition of the user's intention.

Task and annotated language determine the general, event, or translation rubric; T8 uses the assistance rubric. The main comparison and paired controls require a valid Correct or Incorrect verdict.

Failed calls and malformed verdicts remain unresolved and are not converted to model errors. The separate T5 and modality diagnostics retain their earlier parser, which treats strings containing "incorrect" and persistent API exceptions as failures.

The judge assesses compatibility with the textual reference, without independently verifying scene facts absent from it. Equivalent content in another language is accepted, so language-group results do not measure language compliance. The exact assistance prompt appears in Appendix C.3; verdict-processing records accompany the reproducibility materials.

Incomplete interactions. Each Gemini configuration returns complete traces for 1,080 of 1,083 interactions. The remaining three interactions each contain one target, in T1, T3, and T6. These targets count as unsuccessful, retaining all 1,336 targets in the evaluation.

B.4 First-response assignment for T1–T7

All T1–T7 targets are point annotations. For the six-task main comparison, streaming outputs first receive their onset timestamps, with initialization responses removed as specified in Appendix B. LiveCC chunks within two seconds of a group's first post-query chunk are assembled before assignment. Trailing continuation markers are removed and immediately repeated identical chunks within a group are counted once. Repeats in later groups remain separate responses. A group is not extended beyond two seconds to complete its meaning, and it is not discarded simply because its last chunk has a continuation marker. This avoids interpreting prolonged commentary as silence.

Outputs are then sorted by timestamp and merged before assignment: keep the first output, discard outputs within two seconds of its timestamp, then repeat from the next remaining output. Each group retains only its first text and timestamp. Grouping is anchored to the retained output, not chained through consecutive gaps.

For one interaction, order target times as $\tau_1 \leq \cdots \leq \tau_K$ and let U_k contain outputs consumed by earlier targets, initially empty. Define

$$b_1 = -\infty, \qquad b_k = \frac{\tau_{k-1} + \tau_k}{2} \ (k > 1), \qquad J_k = \{j \notin U_k : b_k < t_j \leq \tau_k + 2\}. \tag{3}$$

If J_k is empty, target k fails. Otherwise it consumes the earliest output $j_k \in J_k$, preserving recorded order for timestamp ties, and succeeds if

$$m_k = \mathbb{I}[t_{j_k} \geq \tau_k - 2] C(\hat{y}_{j_k}, y_k). \tag{4}$$

A selected response is consumed even if premature or incorrect, so a later correction cannot rescue that target. The midpoint lower bound is strict and the two-second tolerance endpoints are inclusive. The rule applies to every T7 target, including those after an earlier failed response.

B.5 Ordered matching for assistance

Only temporally eligible pairs under the windows in Appendix B.2 are sent to the judge. For a fixed interaction, let E_{kj} denote the joint content–time eligibility in Eq. 1, omitting the interaction index.

Matching maximizes the number of eligible pairs while preserving target and output order. For the suffix beginning at target k and output j,

$$D_{kj} = \max\{D_{k+1,j},\ D_{k,j+1},\ E_{kj}(1 + D_{k+1,j+1})\}, \tag{5}$$

with zero-valued boundaries. An eligible diagonal is preferred when it achieves the optimum. Otherwise, a target is skipped only if this gives a strictly higher suffix value than skipping an output. One output can cover at most one target. Unlike T1–T7, a later eligible response can match after an earlier failed attempt. Target recall and complete-sequence coverage then follow Eqs. 2 and 6.

Unmatched responses incur no penalty. Appending outputs while retaining the existing outputs and their order cannot reduce the optimal number of matches. Timely reference-target recall therefore measures coverage, without measuring false-alarm rate or intervention cost. References do not enumerate every useful intervention, so an unmatched response cannot automatically be labeled invalid. We therefore report output volume alongside coverage in Appendix D.2.2.

For LiveCC, continuation chunks ending in an ellipsis are concatenated until the next completed chunk and credited at that last chunk's media time. An unfinished final fragment remains at its last available time and is marked incomplete. Other adapters retain their own output boundaries. This processing uses no target information and never credits later content at utterance onset.

B.6 Task aggregates and complete-sequence coverage

Let $m_{ik} \in \{0, 1\}$ indicate target success under the applicable rule, and let $\mathcal{S}$ be the set of interactions being reported. The target score $S(\mathcal{S})$ in Eq. 2 divides total successes by total targets. We report this as task accuracy for T1–T7 and target recall for T8. Table 2's Avg. equally weights the six task accuracies for T1–T4, T6, and T7. Table 3's Overall pools all 361 targets across T8-A, T8-B, and T8-C rather than averaging the three subtask recalls. T5 is retained as a separate rolling diagnostic and marked N/A for prompted inputs, which omit an explicit elapsed-time clock. In T7 and T8-C, an interaction with two successful targets out of three contributes two successes and three targets to this score.

For successive guidance in T8-C, we also report the fraction of interactions in which every reference step is covered:

$$C_{\text{sequence}}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbb{I}\left[\sum_{k=1}^{K_i} m_{ik} = K_i\right]. \tag{6}$$

Table 3 reports the complete-sequence numerator as Complete T8-C, out of 26 interactions. This distinguishes isolated correct steps from coverage of the entire annotated sequence; it does not measure closed-loop task completion.

B.7 Paired uncertainty estimates

Unless otherwise stated, reported differences use 95% percentile confidence intervals from 10,000 bootstrap resamples of video-record clusters. Each resample retains all targets and paired conditions within a selected record. We recompute the stated aggregate, preserving task weighting where applicable. Differences are in percentage points. These exploratory intervals have no multiplicity correction and quantify sampling uncertainty, excluding judge error and API run-to-run variation. A zero-width interval reflects constant observed paired differences, not population equivalence. The complete-guidance permutation reference in Appendix D.3 is defined separately.

B.8 Worked examples of response assignment

Figure 5 contrasts the two assignment rules using constructed examples. Content verdicts are stipulated for illustration; Appendix E shows real model outputs and stored judgments.

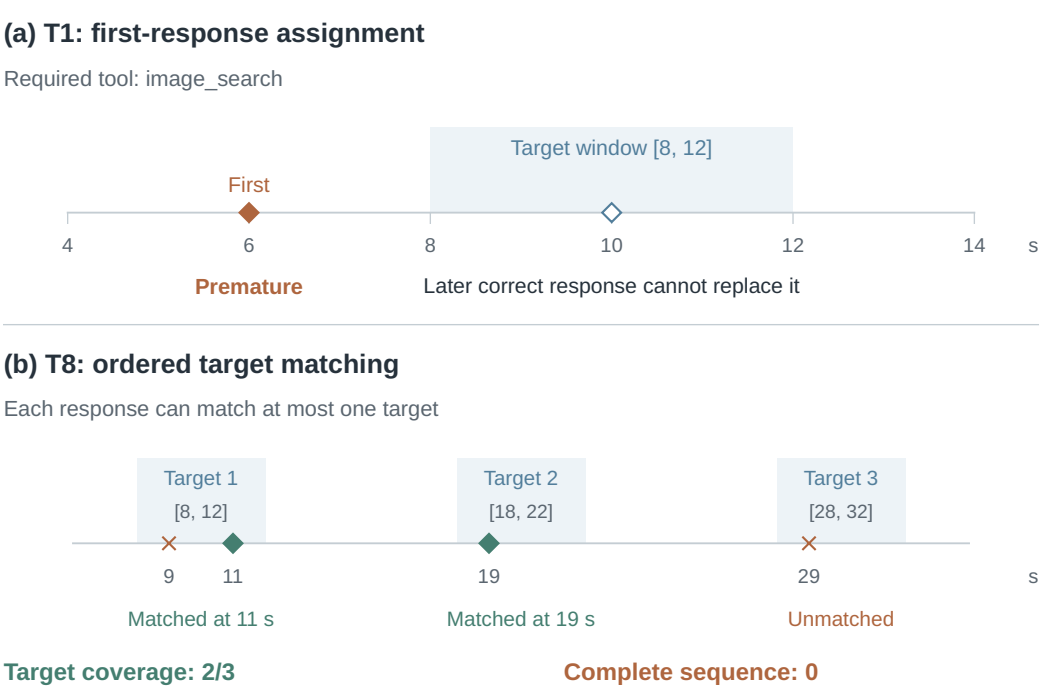


Figure 5: **Timing, content, and assignment play separate roles.** (a) T1 assigns the first response, even when a later response would be eligible. The 6- and 10-second outputs are in separate groups because they are more than two seconds apart. (b) T8 can match a later eligible response after an incorrect attempt. Shaded intervals are inclusive valid windows. Diamonds mark correct content and crosses mark incorrect content; a correct response can still fail timing or remain unassigned. These are constructed scoring examples.

First-response assignment (T1–T7). The target requires `image_search` at 10 seconds. The model emits that tool at 6 seconds and again at 10 seconds. The first response precedes the $[8, 12]$ window and determines failure, giving $m_{i1} = 0$; the later response cannot replace it. Conversely, a timely `web_search` would fail tool identity. Both content and timing are necessary.

Ordered matching (T8). The first target matches at 11 seconds after an incorrect 9-second attempt, and the second matches at 19 seconds. The third has only an incorrect candidate, giving $2/3$ target coverage and zero complete sequences. If an eligible response at 31 seconds also covers the third target, coverage becomes $3/3$ and the sequence is complete. Each response can match at most one target. Unmatched outputs do not reduce recall and are not automatically false positives, which is why coverage is reported alongside output volume.

C Experimental Details

C.1 Input modalities and context

At decision time t, prompted models receive user context and video frames sampled from the available prefix $[0, t]$ at 1 fps, capped at 32 frames. Frame presentation timestamps never exceed t. Native scene audio uses the corresponding available prefix or, for streaming adapters, successive causal chunks. V+ASR uses only completed transcript segments under the declared audio-dependency and context policy. It does not necessarily transcribe the whole prefix or represent non-speech sounds. Query attachments follow the supported input path. Native adapters retain their own context windows.

All T8 configurations receive annotated query text or the general assistance instruction for T8-A. Table 6 distinguishes scene audio from a separate spoken-query attachment. V+ASR assistance uses completed independent one-second audio chunks. All 85 T8 spoken-query attachments belong to

Table 6: Assistance input paths. "Text" in the last column means the annotated query text without consuming the separate query-audio attachment. Every configuration also receives the annotated text.

Model	Interface	Scene audio	Spoken-query attachment
Proprietary Models			
Gemini-3.1-Pro	Prefix	Native prefix	Audio and text
Gemini-3.5-Flash	Prefix	Native prefix	Audio and text
Qwen3.5-Omni-Plus	Prefix	Native prefix	Audio and text
Open-source MLLMs			
Qwen3-Omni	Prefix	Native prefix	Audio and text
Qwen2.5-Omni	Prefix	Native prefix	Audio and text
Qwen3-VL	Prefix	Completed-chunk ASR	Text
InternVL3.5	Prefix	Completed-chunk ASR	Text
Streaming Models			
LiveCC	Stateful	None	Text
MMDuet2	Stateful	None	Text
MOSS-VL	Stateful	None	Text
MiniCPM-o 4.5	Stateful	Native chunks	Text
AURA	Rolling video	None	Text

the Chinese QA group; English queries use translated text with unchanged scene audio, so language-group differences do not isolate language alone.

Short audio chunks are padded with silence where required by the model's processor, without reading future audio or advancing the observation clock. Adapter-specific thresholds and causal-audio execution records are included in the reproducibility materials.

AURA adaptation. Our AURA (Lu et al., 2026a) adapter uses the official checkpoint with up to 30 recent one-second video segments (151,200 pixels per frame), reattaches the request at each decision, and uses silent placeholders for earlier assistant turns. It receives no raw scene audio and uses annotated query text for assistance, cached ASR for historical audio-only queries, and original image attachments.

C.2 Model sizes and generation settings

MiniCPM-o 4.5 and MOSS-VL have 9B and 11.3B full-model parameters, respectively, despite both using a Qwen3-8B language backbone. Proprietary parameter counts are undisclosed.

Local prompted models use greedy decoding; hosted calls use temperature zero. Gemini-3.1-Pro and Gemini-3.5-Flash retain provider-default thinking, so their limits are not answer-only token budgets. MiniCPM-o 4.5 uses greedy duplex decoding, 20 spoken tokens per chunk, and a 30-unit context window. LiveCC uses a repetition penalty of 1.05 and resets at 28,000 tokens; MMDuet2 resets at 100,000. MOSS-VL uses greedy session generation with a 512-token setting and a pace of 12 tokens per second. AURA uses temperature zero and a 512-token limit. Context retention and computation are not equalized across interfaces. Recorded model revisions and configuration fingerprints are retained in the reproducibility materials.

C.3 Core prediction and scoring prompts

The excerpts below document the instructions that determine when to respond and what constitutes target coverage. Model-specific media serialization is described above. Prompt wording is quoted from the recorded implementation; bracketed field descriptions below denote instance-specific input. The metadata above each divider describes usage and is not part of the quoted instruction.

Explicit-request decisions. Prompted T1–T7 adapters prepend the following instruction before the standing request and the disclosed tool block:

T1–T7: response decision

Use: Prompted decisions under an explicit request.

Inputs: Causal scene prefix, standing request, and public tool definitions.

Output: ABSTAIN, a response, or an inline tool call.

You are a proactive, always-on streaming video assistant. You are shown video frames sampled up to the CURRENT moment, plus the user's standing request. Decide whether now is the right moment to respond.

- If the event/condition the user asked about has NOT happened yet, reply with exactly: ABSTAIN

- If it HAS happened by now, give the appropriate response.

The tool block supplies public names and definitions, the inline `<call:NAME>` schema, and the instruction to use plain text when no tool is warranted. It discloses neither target labels nor tool frequencies. Streaming adapters retain their native standby mechanism while receiving the tool disclosure. The request follows as `[Standing request]` and its text.

Assistance context and response policy. T8 supplies a user-profile sentence: either "The user cannot see the scene. Convey relevant visual information in words." or "The user cannot hear the scene. Convey relevant audible events and speech in text; use the original audio and supporting visible clues." The requested language is then specified, followed by:

T8: shared grounding instruction

Use: Shared evidence constraint for T8-A/B/C.

Inputs: Current causal scene evidence and the user profile/request.

Output: Grounded assistance when a response is warranted.

Use only the provided scene evidence up to now. Be accurate, concise and actionable. Do not guess unseen objects, text, directions, distances or unheard sounds.

The subtask-specific sentence is:

T8-A/B/C: task instructions

Use: Select the instruction for the evaluated assistance subtask.

Inputs: User profile and scene evidence; T8-B/C also include the request.

Output: A reminder, an answer, or the currently justified guidance step.

T8-A: Monitor for useful assistance and timely safety reminders. State what is happening and an appropriate action. Do not wait for a question. Avoid redundant reminders.

T8-B: Answer the user's question once the evidence is sufficient.

T8-C: Guide the user's intention with short steps as the observed state changes. Correct visible mistakes. Give only the step currently justified, not future steps.

Autonomous prompted calls end with "If no useful new response is warranted now, output exactly ABSTAIN." Native streaming uses "Stay silent when there is no useful new assistance to provide." T8-B/C append the user request; T8-A does not append an event-specific query. In the paired event-guidance controls, a reviewed event instruction is appended. Under either guidance condition, scheduled calls request an answer at an annotated valid time with the sentence:

Scheduled response condition

Use: Scheduled arms of the paired T8-A guidance control.

Inputs: Current evidence at an annotated response opportunity.

Output: One assistance response; the call supplies timing, not reference content.

An external scheduler requests an answer now. Give the most useful assistance supported by the current evidence. The scheduler does not reveal an event or answer.

This wording describes the model-facing input; the experiment still supplies temporal information through the scheduled call itself.

Assistance content judge. The recorded scoring prompt below evaluates reference-target coverage.

Assistance content judge

Use: Target-specific content scoring for T8.

Inputs: User profile/request, reference target, and prediction; no media or time.

Output: Exactly Correct or Incorrect; the scorer handles time and assignment.

Judge the content of one accessibility assistance response.

The reference is an annotated target, not the only valid thing to say in the scene.

Correct means the prediction communicates this same target information and any essential action without contradictory directions, unsafe advice, incorrect facts or invented critical details.

Equivalent wording is fine. A generic warning alone does not cover a specific target. A C step must provide the reference's currently actionable guidance, not just repeat the user's intention.

Ignore timing (scored separately). Do not penalize a different language here: report language compliance separately. If the prediction concerns another event, mark Incorrect FOR THIS TARGET; that does not mean it is an invalid reminder. Do not infer validity of unannotated events.

User profile/request: {question}

Reference target: {gt}

Prediction: {pred}

Return exactly Correct or Incorrect.

The prompt's instruction about language does not add a language-compliance metric to the reported tables. Its verdict concerns semantic correspondence to the supplied target; temporal eligibility and ordered assignment are handled separately by the scorer. General, event, and translation criteria are summarized in Table 5.

D Additional Results and Diagnostics

The analyses below expand the three main questions: timely action, query-free initiative, and sustained coverage. Evaluation sensitivity and failure analyses follow these results. Paired confidence intervals (CIs) follow Appendix B.7. Complete auxiliary grids are provided in the accompanying `supplementary_data/` directory, with an index, machine-readable cells, and the original result tables.

D.1 Timely action: tool choice and event timing

Table 7 separates on-time first responses, joint tool-invocation accuracy, and tool accuracy conditional on timing. If T denotes an on-time first response and U the correct tool set, then $P(T \cap U) = P(T)P(U \mid T)$. Joint and on-time rates use all 209 targets; conditional accuracy uses each model's own timely subset. Per-tool scores retain 125 image-search, 72 web-search, and 12 audio-search targets.

The full table also reveals failures in tool selection among timely streaming responses: MOSS-VL selects the correct set in three of 35 timely responses, whereas LiveCC has 81 timely responses and no joint successes. These are outcomes of the evaluated model–adapter configurations.

D.1.1 Request and timing priors

The request-only control supplies the same 209 request texts and public tool definitions to Gemini-3.5-Flash, Qwen3-VL, and Qwen2.5-Omni, excluding scene media, attachments, target times, answers, and interaction identifiers. It asks for image search, web search, audio search, or abstention if the text is insufficient. All models use a 2,048-token limit, with temperature zero for Gemini-3.5-Flash and greedy local decoding. Exact-label scoring counts abstention or invalid labels as incorrect.

Table 7: T1 timing and tool accuracy (%). Joint and on-time scores use 209 targets; tool|time uses each model's timely responses. Per-tool denominators are 125 image, 72 web, and 12 audio targets. Incomplete targets remain in the denominator.

Model	On-time	Joint	Tool|time	Image	Web	Audio
Gemini-3.1-Pro	36.84	22.01	59.74	18.40	25.00	41.67
Gemini-3.5-Flash	36.36	28.71	78.95	20.80	41.67	33.33
Qwen3.5-Omni-Plus	37.80	31.10	82.28	25.60	38.89	41.67
Qwen3-Omni	35.89	25.84	72.00	17.60	37.50	41.67
Qwen2.5-Omni	26.79	10.53	39.29	7.20	15.28	16.67
Qwen3-VL	33.01	19.14	57.97	12.00	30.56	25.00
InternVL3.5	36.84	28.23	76.62	22.40	38.89	25.00
LiveCC	38.76	0.00	0.00	0.00	0.00	0.00
MiniCPM-o 4.5	5.74	1.44	25.00	0.00	2.78	8.33
MMDuet2	13.40	0.00	0.00	0.00	0.00	0.00
MOSS-VL	16.75	1.44	8.57	0.80	2.78	0.00
AURA	36.36	0.00	0.00	0.00	0.00	0.00

Gemini-3.5-Flash recovers 202/209 tool labels from the request alone (Table 8a); this does not identify the search referent in the stream.

We then emit either the majority tool (image search) or each model's frozen request-only choice at four fixed delays after query arrival. Policies use only query time and clip duration, withholding abstentions and calls beyond clip end. Table 8b reports all four delays with the same 209-target denominator and two-second scoring tolerance. Gemini-3.5-Flash reaches 67.46% at three seconds, the descriptive maximum among these delays. No independent development set selected this delay, and the closed-label instruction and response opportunities differ from the main evaluation. These controls diagnose request and timing priors rather than provide a matched model ranking.

Table 8: Request and timing priors on 209 T1 targets, without scene input. (a) Request-only tool accuracy and class recalls (%); class sizes are 125/72/12. Mean recall weights classes equally, and abstentions count as errors. (b) Joint hits (%) when each tool policy is emitted at a fixed delay after query arrival. All tested delays are shown; their maximum was not independently selected.

(a) Tool choice from the request

Predictor	Accuracy	Mean recall	Image	Web	Audio	Abstain
Always image search	59.81	33.33	100.00	0.00	0.00	0
Gemini-3.5-Flash	96.65	97.54	96.80	95.83	100.00	4
Qwen3-VL	50.72	64.80	20.80	98.61	75.00	24
Qwen2.5-Omni	61.72	68.43	40.00	98.61	66.67	0

(b) Tool choice plus a fixed response delay

Tool-choice policy	+0 s	+1 s	+3 s	+5 s
Always image search	40 (19.14)	67 (32.06)	86 (41.15)	48 (22.97)
Gemini-3.5-Flash	73 (34.93)	108 (51.67)	141 (67.46)	77 (36.84)
Qwen3-VL	44 (21.05)	57 (27.27)	74 (35.41)	40 (19.14)
Qwen2.5-Omni	50 (23.92)	71 (33.97)	92 (44.02)	48 (22.97)

D.1.2 PAIRED SCENE ABLATION

Gemini-3.5-Flash is rerun on all 209 T1 interactions with and without scene input. Scene removal excludes video, audio, and transcripts while retaining the request, its attachments, and tool definitions. The decision grid, prefix sampling, decoding, and first-response scorer are identical. Execution stops at the first response, which determines each single-target score; nonresponding runs cover the full grid. These stopped traces do not measure full-clip output volume.

Scene-present attains 60 joint hits and 77 on-time responses, compared with 67 and 72 for scene-removed. Both have 131 premature responses, with one versus six late or missing responses. Paired successes are 58 shared, two scene-present only, and nine scene-removed only. The joint-score difference is -3.35 points (95% CI: $[-6.70, -0.48]$, 209 clusters). Both arms are new runs; the main Gemini-3.5-Flash row has 60 hits but 76 on-time responses and one incomplete interaction. The ablation measures input-removal effects, without establishing counterfactual tracking of event occurrence or timing.

D.2 Query-free initiative: reminder coverage and response policy

D.2.1 Event guidance and scheduled responses

The paired experiment uses the same 122 T8-A episodes for Gemini-3.5-Flash, Qwen3-Omni, and Qwen2.5-Omni. It crosses event-specific (E) or general (G) instructions with scheduled (S) or autonomous (A) responding. Scheduled conditions request one response at an annotated valid time. Autonomous conditions use the full causal decision grid without previous responses in the prompt.

For episode i, let $Y_i(g, s)$ indicate a timely target match. Event-guidance effects are $\Delta_s = 122^{-1} \sum_i [Y_i(E, s) - Y_i(G, s)]$; the scheduling interaction is $\Delta_S - \Delta_A$. Table 9 places the four condition results alongside guidance effects and their interaction, using 114 video-record clusters. Only Qwen2.5-Omni shows a statistically supported guidance-by-scheduling interaction. Three target windows contain no autonomous decision time. Restricting to the 119 reachable episodes (111 clusters) preserves every hit count except Qwen2.5-Omni's event-scheduled count, which falls from eight to seven. The reachable-subset interaction is $+6.72$ points for Qwen2.5-Omni (95% CI: $[+1.71, +12.00]$). Gemini-3.5-Flash's interval is $[-2.52, +16.10]$ and Qwen3-Omni's is $[-4.17, +7.32]$, both including zero. The full contrast matrix, including the two scheduling effects and both episode sets, accompanies the supplementary data. Scheduling changes temporal information and response opportunities together, so it does not isolate a planning mechanism.

Table 9: Event guidance and response scheduling on the same 122 T8-A episodes. (a) Hits (recall, %) in the four conditions. Scheduled (S) supplies an annotated response opportunity; Autonomous (A) uses the full causal grid. (b) Event-minus-general effects and their S-minus-A interaction in percentage points, with 95% paired intervals from 114 video-record clusters.

(a) Condition results

	Event		General	
Model	**Scheduled**	**Autonomous**	**Scheduled**	**Autonomous**
Gemini-3.5-Flash	30 (24.59)	30 (24.59)	19 (15.57)	27 (22.13)
Qwen2.5-Omni	8 (6.56)	2 (1.64)	0 (0.00)	3 (2.46)
Qwen3-Omni	8 (6.56)	5 (4.10)	4 (3.28)	3 (2.46)

(b) Paired guidance effects [95% CI]

Model	**Event effect, S**	**Event effect, A**	**Interaction**
Gemini-3.5-Flash	+9.02 [+0.82, +17.36]	+2.46 [-5.83, +11.21]	+6.56 [-2.52, +15.97]
Qwen2.5-Omni	+6.56 [+2.48, +11.29]	-0.82 [-4.27, +2.52]	+7.38 [+2.46, +12.80]
Qwen3-Omni	+3.28 [-1.65, +8.26]	+1.64 [-2.42, +5.83]	+1.64 [-4.07, +7.20]

D.2.2 Response history and response cooldown

Three new Gemini-3.5-Flash runs use all 122 episodes under general assistance instructions: no history, the five most recent timestamped responses, and the same history with a 10-second cooldown between responses, measured in media time. Each arm receives only its own actual response history, without reference answers or match feedback. Other inputs, decision times, and decoding settings are fixed. The model is informed of the cooldown rule. After a response, an external controller skips generation calls for ten seconds. Later decisions receive the current causal prefix. Abstention does

not start a cooldown. The intervention changes both output opportunities and call count, rather than holding compute fixed.

Table 10: Gemini-3.5-Flash history and generation-time cooldown controls on 122 T8-A episodes. All arms are regenerated. History contains the five most recent timestamped responses; cooldown enforces ten seconds between outputs. Paired intervals use 114 video-record clusters.

Condition	**Hits (recall, %)**	**Outputs**	**Outputs/min**
No history	22 (18.03)	7,833	58.39
History	22 (18.03)	3,283	24.47
History + cooldown	10 (8.20)	808	6.02
Paired recall changes (percentage points)			
Contrast		**Difference**	**95% CI**
History − no history		$+0.00$	$[-8.13, +8.00]$
History + cooldown − history		-9.84	$[-17.24, -2.46]$

History reduces outputs by 58.09% with the same aggregate coverage of 22/122, but only ten targets succeed in both arms, with 12 gained and 12 lost. The difference interval does not establish equivalence. Adding cooldown reduces coverage to 10/122 and outputs to 808.

These three conditions do not estimate a history-by-cooldown interaction because a no-history-plus-cooldown arm is absent. Reduced volume alone also does not establish reduced semantic redundancy.

These runs are distinct from Table 3 and the guidance experiment: the regenerated baseline has 22 hits and 7,833 outputs, compared with the earlier Gemini-3.5-Flash general condition's 27 hits and 7,813 outputs. All contrasts in Table 10 use the regenerated arms and 114 paired video-record clusters. These findings concern Gemini-3.5-Flash under the specified policies.

D.2.3 Filtering recorded outputs

We also apply first-K and minimum-spacing filters to the earlier autonomous trajectories. First-K retains the first 1, 3, 5, or 10 responses per episode; spacing filters retain the first response and subsequent responses at least 5 or 10 seconds after the last retained one. Filtering uses no reference times, and matching is recomputed on retained outputs. The complete model-by-policy grid is provided in the supplementary data.

Under general instructions, five-second spacing leaves Gemini-3.5-Flash with 11 of 27 target matches and 1,642 of 7,813 outputs. For Qwen3-Omni, the same spacing leaves two of three matches and 1,666 of 8,213 outputs. Unlike the response cooldown, these filters cannot change subsequent model behavior. Both analyses report coverage together with output volume; unmatched outputs are not automatically false positives (Appendix B).

D.3 Sustained coverage: steps and complete guidance

The 26 T8-C interactions contain 20 three-step sequences, four four-step sequences, and one each with five and six steps. Table 3 reports overall step and complete coverage. We examine the largest length group below; the full model-by-length matrix accompanies the supplementary data. The four-, five-, and six-step groups are too small to establish a length effect.

To interpret complete coverage, we independently permute the binary step-success outcomes across episodes of the same length within each step position. For n episodes of length K and marginal success rates p_k, the expected complete count is

$$\mathbb{E}_{\text{perm}}[N_{\text{complete}}] = n \prod_{k=1}^{K} p_k. \tag{7}$$

Figure 6 shows all models on the 20 three-step episodes, with 95% reference intervals from 10,000 permutations (seed 20260921). These intervals describe the permutation distribution, not uncertainty

in the observed rate. The reference assumes exchangeability across episodes within length and does not control shared episode difficulty.

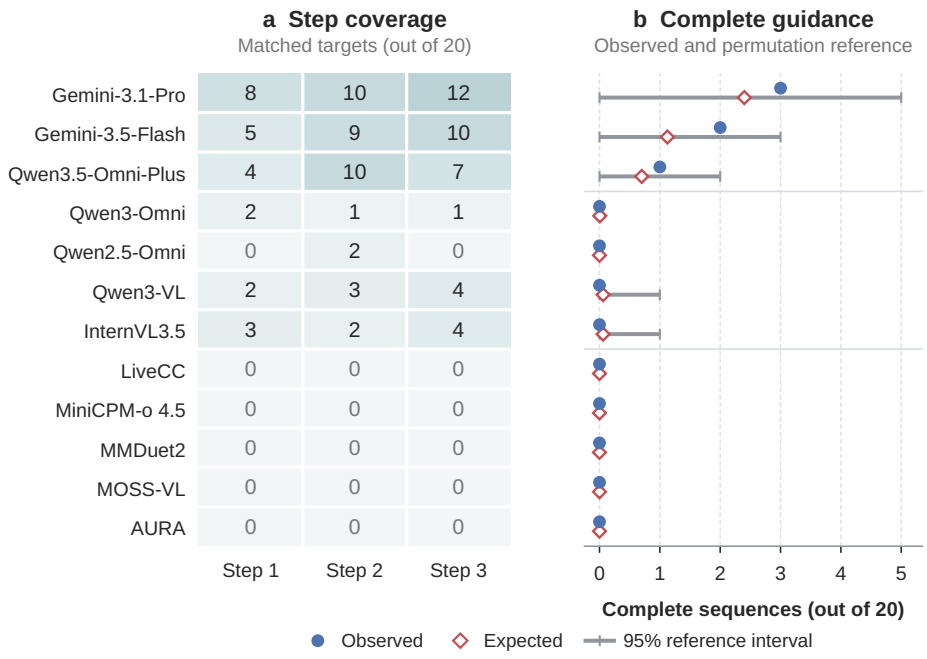


Figure 6: **Step coverage and complete guidance on 20 three-step episodes.** (a) Matched targets at each step position, with a denominator of 20 per cell. (b) Observed complete counts, permutation expectations, and central 95% reference intervals. Each permutation preserves the marginal counts in (a). The reference does not model episode difficulty or identify a memory mechanism.

For Gemini-3.1-Pro, the step counts are 8/20, 10/20, and 12/20, totaling 30/60 covered steps. Its expected complete count is $20 \times 0.4 \times 0.5 \times 0.6 = 2.40$, with a reference interval of $[0, 5]$ containing the observed three completions. Gemini-3.5-Flash's observed two completions likewise lie within $[0, 3]$ around an expectation of 1.125. These results motivate reporting complete guidance alongside step coverage, without establishing an additional continuity deficit beyond the step-level outcomes. Both measures concern the recorded reference trajectory, not closed-loop task success.

D.4 Evaluation sensitivity and failure analysis

D.4.1 Decision schedules

A cohort selected before inference contains 20 interactions from each of T1–T4, T6, and T7: 120 interactions, 143 targets, and 118 video-record clusters. Gemini-3.5-Flash and Qwen3-VL each run two schedules with the same query preparation, historical context, frame sampling, modality, decoding, and judge. Target-local probing uses one-second decisions within four seconds of each reference time, clipped to query and media boundaries. Continuous probing uses the main query-plus-integer grid. Thus both the extent and phase of the schedule change.

Table 11 reports equally weighted six-task means. Gemini-3.5-Flash's accuracy falls by 7.50 points as premature responses rise from 58.74% to 67.13% of targets. First-response scoring exposes early actions that target-local probing cannot observe. Qwen3-VL's change has an interval spanning zero. Premature fractions weight targets equally; these within-model contrasts do not isolate architecture or modality effects.

Table 11: Decision-schedule effects on 120 paired interactions (143 targets). Accuracy equally weights T1–T4, T6, and T7. Differences are continuous minus target-local, in percentage points, with 95% paired intervals.

Model	Target-local	Continuous	Difference [95% CI]
Gemini-3.5-Flash	26.49	18.99	-7.50 $[-13.77, -1.64]$
Qwen3-VL	4.11	3.28	-0.83 $[-3.88, +2.11]$

D.4.2 Temporal-tolerance sensitivity

We rescore the same 418 trajectories at one-, two-, and four-second tolerances, keeping predictions, tool extraction, first-response assignment, and the separate two-second output-grouping rule fixed. Table 12 shows that absolute scores change substantially. All scene-present minus scene-removed estimates are negative, but the one-second interval includes zero. The absence of a positive point estimate persists across these windows; neither score magnitude nor statistical support is invariant. The main tolerance remains two seconds, and this sensitivity result is specific to the Gemini-3.5-Flash control.

Table 12: Temporal-tolerance sensitivity of the Gemini-3.5-Flash scene control. Cells give joint accuracy (%) and hits out of 209. Differences are scene present minus removed, in percentage points, with 95% paired intervals. The output-grouping rule remains fixed at two seconds.

Tolerance (s)	Scene present	Scene removed	Difference [95% CI]
1	10.53 (22/209)	11.96 (25/209)	-1.44 $[-3.83, +0.96]$
2 (primary)	28.71 (60/209)	32.06 (67/209)	-3.35 $[-6.70, -0.48]$
4	51.67 (108/209)	64.11 (134/209)	-12.44 $[-17.70, -7.18]$

D.4.3 Timing and content failures

Table 13 decomposes the main results for state tracking (T4) and continuous responses (T7). Gemini-3.5-Flash's T4 failures are predominantly premature (176/210 targets), while T7 has both premature responses (234/462) and timely content failures (156/462). A premature response fails regardless of its content, whose correctness cannot be inferred from this breakdown.

Table 13: First-response outcomes on state tracking (T4) and continuous responses (T7), using the main-table trajectories. Counts sum to the target denominator in each row. Content failure denotes an on-time response with incorrect content.

Model	Task	Targets	Success	Premature	Content fail	Late/missing
Gemini-3.1-Pro	T4	210	23	182	5	0
Gemini-3.1-Pro	T7	462	51	230	149	32
Gemini-3.5-Flash	T4	210	21	176	10	3
Gemini-3.5-Flash	T7	462	52	234	156	20

D.4.4 Scene audio with the request held fixed

Excluding query-audio attachments and audio-type queries leaves 1,153 interactions and 1,406 targets. Conditions share query text, target-local decisions, and context policy within each model. V provides video, V+ASR adds transcripts, and V,A adds native scene audio. These runs use a separate schedule and earlier verdict handling (Appendix B) from the main comparison.

Table 14: Scene-audio effects with the query held fixed under target-local probing. (a) Joint hit counts; each model uses the same target denominator. V: video, V+ASR: video and transcripts, V,A: video and native audio. (b) Native-audio contrasts in percentage points with 95% paired intervals (47 auditory and 27 audiovisual video-record clusters). These runs use a separate schedule and earlier verdict handling from the main comparison.

(a) Overall and audio-related target subsets

		Gemini-3.5-Flash			Qwen3-Omni			Qwen2.5-Omni		
Subset	**Targets**	**V**	**V+ASR**	**V,A**	**V**	**V+ASR**	**V,A**	**V**	**V+ASR**	**V,A**
All with T5	1406	316	316	322	112	118	125	39	42	50
Auditory	78	6	6	13	4	8	11	3	5	5
Audiovisual	35	4	4	5	0	2	1	0	1	0

Paired scene-audio effects (percentage points)

Trigger	**Comparison**	**Gemini-3.5-Flash**	**Qwen3-Omni**	**Qwen2.5-Omni**
Auditory	V, A − V	$+8.97\,[+3.33,\ +15.71]$	$+8.97\,[+3.41,\ +16.39]$	$+2.56\,[+0.00,\ +7.04]$
	V, A − V+ASR	$+8.97\,[+3.33,\ +15.79]$	$+3.85\,[-2.60,\ +10.98]$	$+0.00\,[-5.19,\ +5.26]$
Audiovisual	V, A − V	$+2.86\,[-9.76,\ +16.13]$	$+2.86\,[+0.00,\ +9.68]$	$+0.00\,[+0.00,\ +0.00]$
	V, A − V+ASR	$+2.86\,[-9.68,\ +16.13]$	$-2.86\,[-13.51,\ +6.45]$	$-2.86\,[-9.09,\ +0.00]$

Table 14 retains the overall result and the auditory and audiovisual subsets. The complete five-way trigger partition and the aggregate excluding T5 (1,291 targets) accompany the supplementary data. These aggregates weight targets equally, unlike the main six-task Avg. Auditory and audiovisual contrasts use 47 and 27 video-record clusters, respectively.

Native audio adds seven hits on 78 auditory targets for both Gemini-3.5-Flash (6 to 13) and Qwen3-Omni (4 to 11), gains of 8.97 points with intervals above zero. Gemini-3.5-Flash gains only 0.43 points overall: aggregation obscures the audio-specific benefit. These effects do not prove audio necessity for every target.

E QUALITATIVE CASES

E.1 TIMELY ACTION AND QUERY-FREE REMINDERS

Figure 7 pairs real source frames with reference requirements and recorded model outputs. These selected cases illustrate failure modes, not their prevalence. Times are rounded for display; verdicts are taken from the stored scorer outputs.

(a) T1 Acting before the requested close-up

USER REQUEST

Automatically look up the person's info the first time a close-up of the guy in purple clothes appears.



28.63 s | 35.70 s | 35.80 s

Valid window [33.80, 37.80]

28.63 s: Q + first call | 35.80 s: reference

Gemini-3.5-Flash 28.63 s

```
<call:web_search(query="Felix Auger-Aliassime")>
```

Stored outcome: premature

(b) T8-A Recommending exit instead of waiting

General visual assistance; no event-specific query



39.49 s

REFERENCE [38.60, 43.00] s

Please note that you have not reached the ground floor yet. Keep waiting.

Gemini-3.1-Pro

39.00 s

The elevator has stopped and the doors are opening. You can exit now.

40.00 s

The elevator doors have opened. You can now exit.

41, 42, 43 s: also recommend leaving

M remains lit

Five in-window candidates, all incorrect

Target coverage: 0/1

Outputs in episode: 57

Figure 7: **Acting early and giving the wrong reminder.** (a) Gemini-3.5-Flash calls a tool at query arrival, before the scored window, and repeats it at 29.00 s. The reference tool is `image_search`; timing already determines failure. (b) The still-lit M button is outlined in the source frame. All five in-window Gemini-3.1-Pro responses recommend exit instead of waiting. Two are quoted in full; other unmatched outputs are not thereby classified as invalid assistance. Neither case establishes an internal cause.

E.2 CODE DIAGNOSIS UNDER CHANGING SCREEN STATES

Figure 8 contrasts locally grounded advice with a later screen-inconsistent diagnosis. Gemini-3.1-Pro identifies a duplicate variable and an extra semicolon, but then claims that the initializer braces have been deleted when both are visible. Useful earlier advice therefore does not establish that subsequent recommendations remain consistent with the current code.

T6 C# code diagnosis as the screen changes

USER REQUEST AT 20.83 s (translated)

I'm going to write and submit code. Keep an eye on me and offer relevant suggestions if I miss a key step or stay stuck on the same problem.

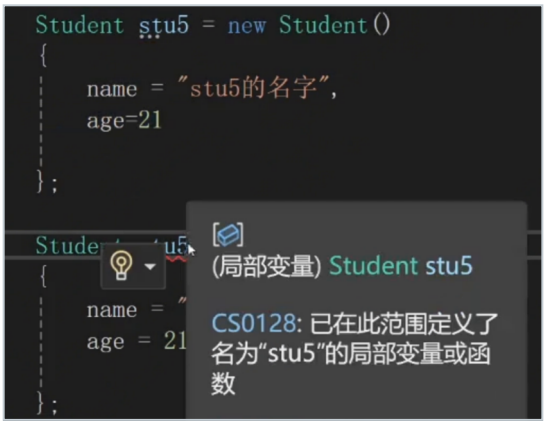
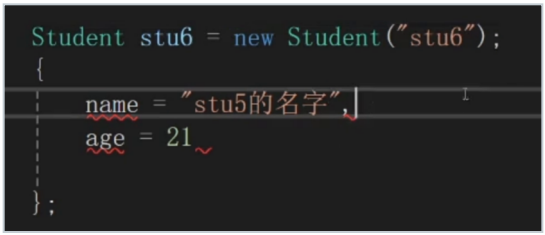
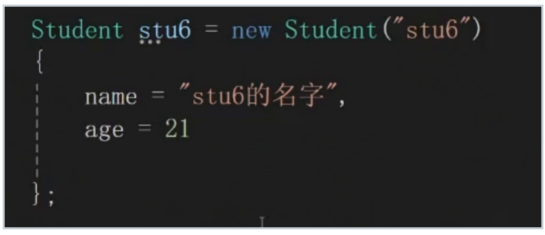
SCREEN DETAIL	Gemini-3.1-Pro (translated excerpts)
20.83 s: duplicate variable	
	You copied the stu5 instantiation code but did not rename the variable after pasting it below. This causes compiler error CS0128: two local variables cannot share a name in the same scope. Rename the second variable to stu6 or another name.
34.00 s: extra semicolon	
	When instantiating stu6, you added an extra semicolon after new Student("stu6"), causing a syntax error in the object initializer that follows. Remove that semicolon [...]
41.00 s: braces already present	
	When instantiating stu6, you deleted the braces {} of the object initializer. [...] The visible braces contradict this diagnosis.

Figure 8: **C# diagnosis across three editing states.** Source-frame crops accompany recorded replies at the indicated times. The request and replies are translated from Chinese; [...] marks omissions. The original target is a reminder to run tests at 60.50 s, against which the stored outcome is premature. The displayed diagnoses illustrate agreement with the screen, not additional scored targets.

E.3 A TUTORIAL REMINDER AT THE REQUESTED STAGE

In Figure 9, the user requests a reminder when a tutorial introduces its second media-import method. Qwen3-Omni responds at 162.00 s, before the reference window. Gemini-3.1-Pro first responds at 191.00 s and receives a stored match. The contrast illustrates why a plausible description of an operation must also be tied to the requested stage of the demonstration.

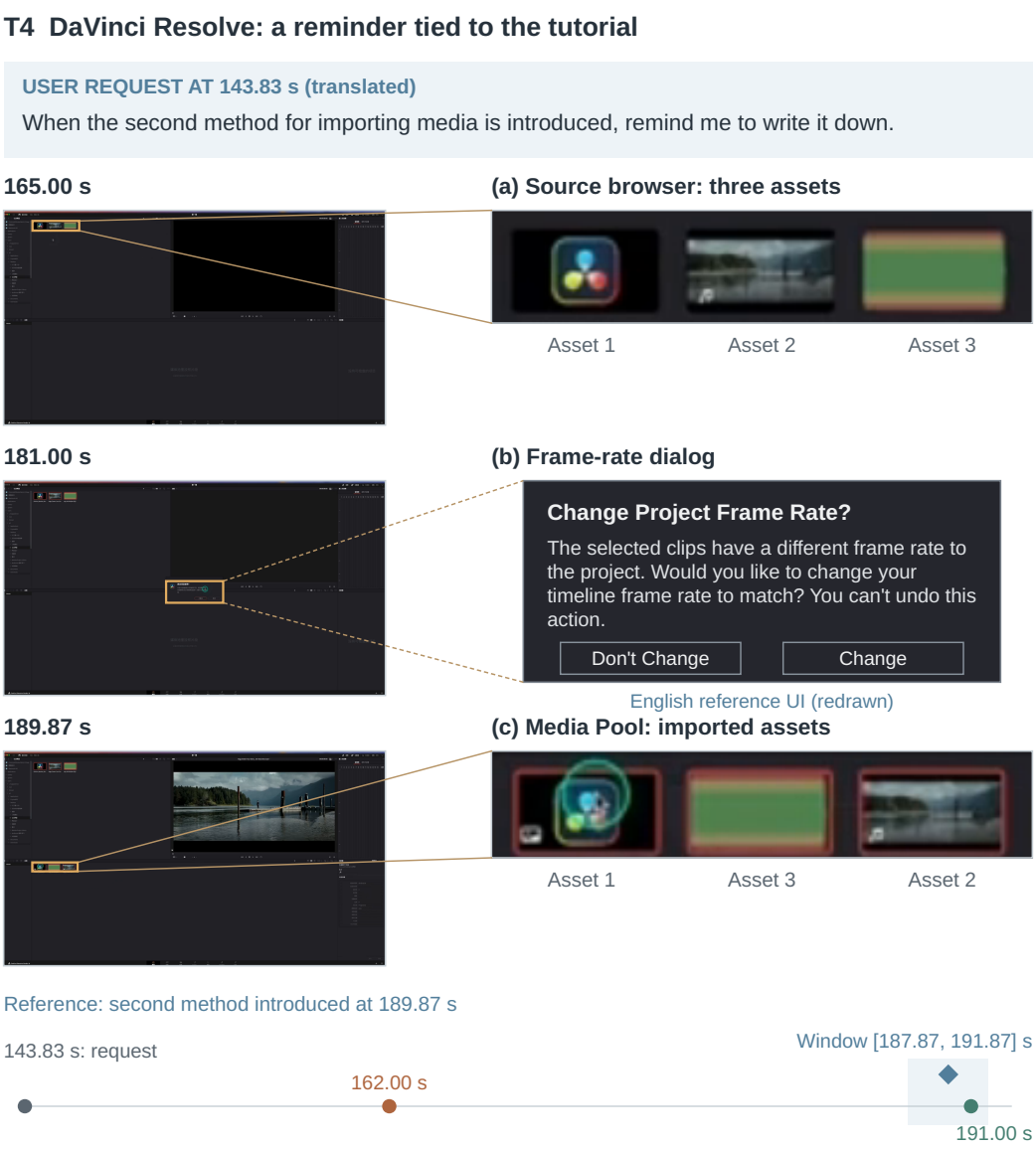


Figure 9: **Following a DaVinci Resolve import workflow.** Original frames appear on the left. Panels (a,c) show source crops with matching asset labels; (b) is a redrawn English reference dialog.[1] The diamond marks the annotated second-method point (189.87 s), with its ±2-second window shaded. Circles mark the request and first replies, all translated in full.

[1]Chris Roberts and Simon Hall, *The Beginner's Guide to DaVinci Resolve 20*, Blackmagic Design, 2025, p. 351. Official PDF.

E.4 Partial coverage within a guidance sequence

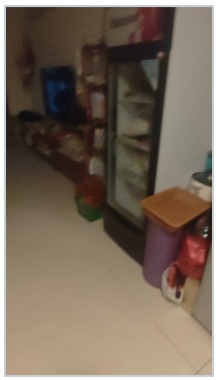
Figure 10 shows two matched steps and a missed final target. All four final-target candidates at 32–35 seconds are incorrect, giving $2/3$ target coverage and zero complete sequences despite 67 outputs. This illustrates partial coverage without identifying a memory or planning mechanism.

T8-C Two covered steps, incomplete guidance

USER REQUEST

I need to take my green water bottle out of the refrigerator.

STEP 1



22.34 s

REFERENCE

[20.35, 24.35] s

Turn to your right. The refrigerator is on your right.

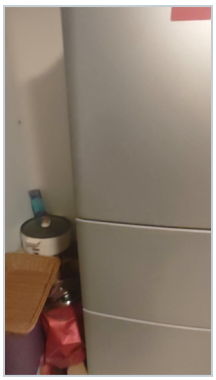
Gemini-3.1-Pro

22.00 s

Turn right to face the refrigerator.

Matched

STEP 2



26.41 s

REFERENCE

[24.42, 28.42] s

The refrigerator is not open. Open the refrigerator door in front of you.

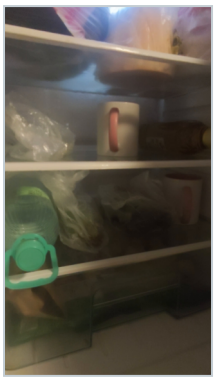
Gemini-3.1-Pro

26.00 s

Open the refrigerator in front of you.

Matched

STEP 3



33.39 s

REFERENCE

[31.40, 35.40] s

The water bottle is on the left of the second shelf from the bottom. Remember to close the refrigerator door after taking it out.

Gemini-3.1-Pro

34.00 s

The green water bottle is on the bottom shelf of the refrigerator, on the left side. Reach down and grab it.

Incorrect

Target coverage: 2/3 **Complete sequence: 0**

Final target: wrong shelf and no instruction to close the door.

Figure 10: **Two covered steps with an incomplete final instruction.** Each column pairs a source state with the full reference and a recorded Gemini-3.1-Pro response. The first two columns show stored matches; the third shows one of four incorrect final-target candidates. The shelf location and closing action distinguish the reference from the prediction. Frame timestamps identify sampled evidence, while output timestamps identify model emissions. An incorrect 21-second attempt does not prevent the later 22-second match under ordered assignment.