OmniInteract: Benchmarking Real-World Streaming Interaction for Real-Time Omnimodal Assistants

Anonymous ACL submission

Abstract

We introduce OmniInteract, a streaming benchmark for real-time omnimodal large language models evaluated through native online inference over audio-visual streams. Unlike offline video understanding or text-prompted streaming QA, OmniInteract preserves the original audio-visual stream and requires models to process it online, without access to future content. User queries and ambient sounds are embedded in the audio track, requiring models to detect multimodal triggers, decide when to respond, and answer while the stream unfolds. OmniInteract contains 250 videos with 1,430 temporally grounded response slots: 1,062 1Q1A slots across real-time, proactive, and nested scenarios, and 368 1QnA slots for continuous task monitoring and step guidance. Each slot includes a trigger, response window, and target answer. We evaluate response correctness, timing, invalid outputs, interruption handling, and context continuity using Interaction-Aware Quality-Timeliness F1, Interruption Diagnostic Suite, and Nested Chain Completion Score. Experiments show that current models remain weak in streaming interaction, with the best overall IA-QTF1 reaching only 0.368 and the best 1QnA IA-QTF1 only 0.052. Further study on mathematical reasoning in full-duplex settings shows that offline capability does not necessarily transfer to online interaction.

1 Introduction

Human–AI interaction is shifting from offline multimodal understanding to continuous, real-time communication (Chen et al., 2025; Zeng et al., 2026; Yang et al., 2025; Liu et al., 2026; Xia et al., 2025; Fu et al., 2025b; Liu et al., 2024). Conventional video-language evaluation typically asks models to answer questions after the relevant content has already been observed (Fu et al., 2025a; Li et al., 2024; Wu et al., 2024), while recent streaming video benchmarks move closer to online perception (Lin et al., 2026b; Niu et al., 2025; Lu et al., 2026b). Meanwhile, omnimodal large language models (LLMs) are integrating vision, audio, speech, and text into unified systems (Chen et al., 2024b,c; Team, 2026; Comanici et al., 2025; AI et al., 2025; Cui et al., 2026). These developments call for an evaluation setting beyond hindsight understanding: a real-time assistant must decide whether to respond, when to respond, and what to say during an ongoing audio-visual interaction.

However, existing benchmarks do not fully capture this coupled decision process. Offline video question answering removes the need to decide response timing by allowing models to access the full video before answering (Fu et al., 2025a; Li et al., 2024; Wu et al., 2024; Hu et al., 2025; Zhao et al., 2025). Most streaming video benchmarks retain temporal inputs, but provide user questions as external textual prompts (Lin et al., 2026b; Niu et al., 2025; Lu et al., 2026b; Wang et al., 2025c,b), bypassing spoken intent recognition from the audio stream. Moreover, existing benchmarks are evaluated on pre-segmented video clips with offline inference, or rely on custom streaming protocols distinct from the models' native real-time inference. As a result, they only partially evaluate the interaction loop required by native real-time assistants: detecting spoken or multimodal triggers, grounding them in visual events and background sounds, responding at the right moment, and avoiding invalid outputs while operating under genuine online streaming constraints. This limitation becomes more evident in full-duplex-oriented scenarios, where users may interrupt, insert new questions, or expect the assistant to resume an unfinished interaction (Défossez et al., 2024; Yao et al., 2025; Lin et al., 2025b,a, 2026a; Cui et al., 2026).

To evaluate this missing interaction loop, we introduce **OmniInteract**, a benchmark that directly evaluates omnimodal LLMs through their native online streaming inference in continuous real-time audio-visual streams. Fig. 1 contrasts this

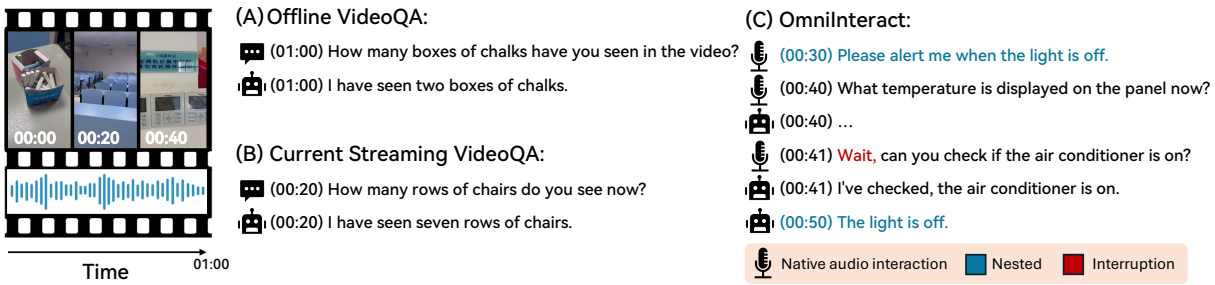


Figure 1: Comparison of offline video QA, text-prompted streaming video QA, and OmniInteract (1Q1A). OmniInteract preserves spoken queries and multimodal events in the original audio-visual stream for timely, interruption-aware, and nested interaction evaluation.

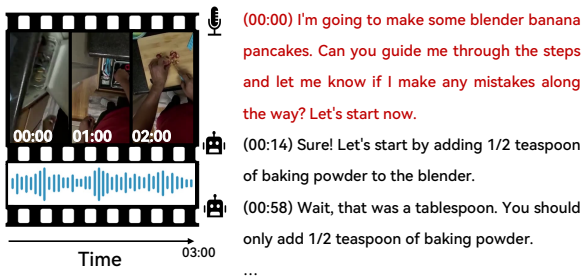


Figure 2: Example of a 1QnA interaction. A single spoken instruction can require multiple temporally grounded response slots as the task unfolds.

setting with offline and text-prompted streaming video QA. Rather than converting interactions into video-text question-answer pairs, OmniInteract preserves them in their native multimodal form: spoken user queries remain in the audio track, while visual events and background sounds remain part of the evolving context. Models must process the stream as it unfolds, without lookahead to future content. This design better reflects real interaction, but it also raises a practical question: how can a continuous audio-visual stream be evaluated when it does not naturally provide fixed question-answer boundaries?

We address this question with an interaction slot formulation. Each slot represents a temporally grounded response opportunity, defined by a trigger, an expected response window, and a target answer. These elements correspond to the three key decisions in real-time interaction: the trigger indicates whether a response opportunity exists, the response window specifies when the model should answer, and the target answer defines what it should say. In this way, the slot formulation makes continuous omnimodal interaction measurable while preserving its temporal and multimodal nature.

Building on this formulation, OmniInteract includes two complementary interaction structures with 250 videos and 1,430 temporally grounded response slots in total. The **1Q1A** split contains 1,062 single-response slots (210 videos), including 638 real-time, 184 proactive, and 240 nested slots. It focuses on localized interactions constructed from self-recorded videos and manual annotations, where each trigger corresponds to one expected answer. The **1QnA** split contains 368 response slots (40 videos) for continuous task monitoring from existing benchmarks, where a single instruction may require multiple temporally grounded responses as the task progresses; Fig. 2 shows a representative example. Together, these splits evaluate whether models can handle both immediate response opportunities and longer-horizon monitoring within the original audio-visual stream.

The slot formulation also guides the evaluation metrics. Since each slot specifies both answer content and a valid response window, answer accuracy alone is insufficient: a semantically correct response may still fail as an interaction if it is produced too early, too late, or outside the intended context. OmniInteract further stresses interaction control with 192 interrupted response slots, including 147 in 1Q1A and 45 in 1QnA, as well as 240 nested slots forming 120 pairs that require models to answer an inserted inner query before resuming the outer query. We therefore propose an **Interaction-Aware Quality-Timeliness F1** (IA-QTF1), together with **Interruption Diagnostic Suite** (IDS) and the **Nested Chain Completion Score** (NCCS), to jointly measure response quality, timing, undesirable outputs, interruption handling, and context resumption.

We evaluate representative omnimodal real-time interaction models on OmniInteract. The results reveal substantial variation across scenarios, with continuous task monitoring remaining the most challenging setting because models must produce multiple temporally grounded responses over an extended stream. We further conduct a focused offline-online comparison on MiniCPM-o 4.5 math-

Table 1: Benchmark comparison. We compare input modalities, query form, online inference, and interaction coverage across prior streaming video benchmarks and OmniInteract.

Benchmark	Modality	Query Modality	Online Inference	1Q1A			1QnA	Interruption
				Real-time	Proactive	Nested		
StreamingBench (Lin et al., 2026b)	V, A	T	✗	✓	✓	✗	✗	✗
OVO-Bench (Niu et al., 2025)	V	T	✗	✓	✓	✗	✗	✗
OmniMMI (Wang et al., 2025c)	V, A	T	✗	✓	✓	✗	✗	✗
ProactiveVideoQA (Wang et al., 2025b)	V, A	T	✗	✗	✓	✗	✗	✗
PhoStream (Lu et al., 2026b)	V, A	T	✓*	✓	✓	✗	✗	✗
OmniInteract (Ours)	**V, A**	**A**	✓	✓	✓	✓	✓	✓

V: Video, **A**: Audio, **T**: Text. *: uses a custom streaming evaluation protocol rather than models' native online streaming inference.

ematical reasoning tasks in a full-duplex-oriented setting (Cui et al., 2026), showing that reasoning quality degrades substantially when the model must reason while simultaneously listening and generating responses. Together, these results highlight a key gap in current omnimodal real-time interaction: strong multimodal understanding or reasoning in offline settings does not necessarily translate into robust real-time interaction.

Our contributions are summarized as follows:

1) We introduce **OmniInteract**, a benchmark for evaluating omnimodal LLMs through their native online streaming inference over continuous real-time audio-visual streams. OmniInteract preserves spoken queries, visual events, and background sounds in the original stream, and covers two complementary interaction structures: **1Q1A** for localized single-response interactions and **1QnA** for continuous task monitoring.

2) We propose an interaction slot formulation that represents each temporally grounded response opportunity with a trigger, an expected response window, and a target answer. Built on this, we develop Interaction-Aware Quality-Timeliness F1, Interruption Diagnostic Suite, and Nested Chain Completion Score, enabling joint evaluation of response content, timing, undesirable outputs, interruption handling, and context resumption.

3) We conduct a systematic benchmark analysis of representative omnimodal real-time interaction models under native spoken-query, online audio-visual interaction, with additional analyses of full-duplex-oriented behaviors. Our results reveal substantial gaps in current models, especially in continuous task monitoring and temporally grounded interaction control.

2 Related Work

2.1 Streaming Video Understanding

Streaming video understanding shifts from offline post-hoc understanding (Fu et al., 2025a; Li et al., 2024; Wu et al., 2024) to real-time online interaction (Lin et al., 2026b; Niu et al., 2025; Lu et al., 2026b; Shen et al., 2026), requiring synchronized perception, decision-making, and response. Recent works address this challenge through temporally aligned long-context modeling (Chen et al., 2024a), streaming token management with compact visual-text windows (Xu et al., 2025), asynchronous perception-decision-reaction pipelines (Qian et al., 2025), proactive response training with dynamic compression (Zhang et al., 2025), multi-turn reinforcement learning for timely responses (Wang et al., 2025a), offline-to-streaming adaptation with memory and activation mechanisms (Wang et al., 2026), and end-to-end continuous observation frameworks (Lu et al., 2026a). These systems make important progress toward online video understanding, but existing benchmarks still only partially capture native real-time interaction. As summarized in Tab. 1, they typically provide user queries as text rather than spoken audio, and evaluate models on pre-segmented clips using offline inference or custom streaming protocols instead of the models' native online streaming inference. These choices decouple response generation from the real-time perception, spoken intent recognition, and timing control required by native streaming assistants.

2.2 Omnimodal Large Language Models

Beyond temporal streaming, omnimodal LLMs extend multimodal interaction by integrating vision, audio, speech, and text within unified systems. Recent models add audio encoders to visual-language backbones (Chen et al., 2024b,c), unify multiple modalities in shared token spaces (Team et al., 2026), scale native audio-visual interaction with mixture-of-experts and speech-generation architectures (Team, 2026; AI et al., 2025), and advance long-context multimodal reasoning over audio-visual inputs (Comanici et al., 2025). These developments enable richer interaction interfaces,

where user intent may appear as speech, background sounds may affect the response context, and visual events may determine when the model should answer. However, evaluation has not fully kept pace with these capabilities. Prior benchmarks cover parts of streaming video understanding, such as real-time or proactive QA, but they generally retain text queries, omit nested or multi-answer interaction structures, and do not evaluate interruption handling under native online inference. OmniInteract targets this gap by combining spoken audio queries, online model execution, 1Q1A and 1QnA interaction structures, and interruption-aware evaluation within the same benchmark.

2.3 Full-Duplex Real-Time Interaction

Streaming video understanding and omnimodal modeling naturally motivate full-duplex real-time interaction, where models process incoming input while generating output for more natural human–AI communication. Early full-duplex studies focus mainly on spoken dialogue, enabling low-latency speech-to-speech interaction without explicit turn segmentation (Défossez et al., 2024) and improving native audio interaction through dedicated training paradigms (Yao et al., 2025). Full-Duplex-Bench evaluates capabilities such as interruption handling, smooth turn-taking, and conversational continuity (Lin et al., 2025b,a, 2026a). At the multimodal level, recent work introduces a time-aligned streaming framework for simultaneous perception, speech generation, and proactive behavior (Cui et al., 2026). These works highlight the importance of interruption handling, overlapping input/output, and context continuation. OmniInteract complements them by evaluating such behaviors in continuous audio-visual streams with temporally grounded spoken-query interactions.

3 OmniInteract Benchmark

3.1 Data Composition

OmniInteract is constructed to evaluate omnimodal LLMs through their native online streaming inference in continuous real-time interaction scenarios. Unlike conventional offline video question answering (Fu et al., 2025a, 2026), where responses are produced after observing a complete video or clip, OmniInteract requires models to process the audio-visual stream as it unfolds, without lookahead to future content. We organize the data around interaction slots, each associated with a trigger, an expected response window, and a target answer (detailed in Sec. 3.3.1). Beyond temporal streaming, OmniInteract further differs from prior streaming video benchmarks that often provide user questions as external textual inputs (Lin et al., 2026b; Niu et al., 2025; Lu et al., 2026b). OmniInteract preserves the original audio-visual stream as the primary interaction context, where user queries are directly recorded in the audio track together with background sounds and visual events. This formulation evaluates whether models can recognize spoken intents, interpret multimodal evidence, and respond at appropriate moments in an end-to-end omnimodal setting.

Table 2: Statistics of OmniInteract. Video counts denote the number of source videos; slot counts denote temporally grounded response slots; interruptions are cross-cutting cases included in the corresponding split.

Split	Type	Videos	Slots	Interruptions
1Q1A	Real-time	210	638	147
	Proactive		184	
	Nested		240	
1QnA	Monitoring	40	368	45
Total		**250**	**1,430**	**192**

Following this formulation, we categorize interaction instances according to whether they require a single response or multiple temporally evolving responses. OmniInteract is therefore organized into two complementary splits: 1Q1A and 1QnA. The 1Q1A split consists of instances where each trigger corresponds to one expected answer, and is further divided into three interaction types. Real-time interaction involves an explicit user query issued during the multimodal stream, where the model is expected to respond immediately based on the available context. Proactive interaction is driven by salient multimodal events rather than an explicit query, requiring the model to continuously monitor the stream and respond only when sufficient evidence or a relevant cue emerges. Nested interaction occurs when a real-time query is inserted within the response window of a proactive interaction, requiring the model to address the inserted query while maintaining the context of the original interaction. The 1QnA split covers cases where a single query or instruction corresponds to multiple valid answers over time. It evaluates whether a model can provide temporally appropriate responses as new evidence appears in the stream, rather than reducing the interaction to one static answer.

Tab. 2 summarizes the resulting split sizes. The

1Q1A split contains 1,062 response slots across real-time, proactive, and nested interactions, while 1QnA contains 368 response slots. The 147 interruptions in 1Q1A and 45 interruptions in 1QnA are annotated as cross-cutting cases within these splits rather than as a separate interaction type.

3.2 Data Curation

Given the different interaction structures of 1Q1A and 1QnA, we adopt different curation strategies for the two splits. Due to the lack of datasets specifically designed for native real-time omnimodal interaction, we curate the 1Q1A split from scratch. We self-record 210 videos in two groups of scenarios. The first group covers daily-life interactions in Chinese, including home activities, gym exercises, museums, shopping, and other common situated interactions (150 videos). The second group covers English mathematical problem-solving, where the user asks questions while the visual stream shows the evolving problem context (60 videos). For real-time interactions, we record explicit spoken queries in the audio track and align each query with the visual evidence needed for answering. For proactive interactions, the user first issues a spoken query whose answer is not yet available; the model must monitor the subsequent audio-visual stream and respond once the required evidence emerges. For nested interactions, we insert a real-time query into the response window of an ongoing proactive interaction, so that the model must answer the inserted query before resuming the original context. For each slot, we manually annotate the trigger, valid response window, and target answer, and verify that the answer is supported by the corresponding audio-visual evidence.

For the 1QnA split, we construct continuous monitoring instances from existing procedural and task-oriented video benchmarks (40 videos), including live step-by-step task guidance (Bhattacharyya et al., 2026; Peddi et al., 2024) and egocentric error detection (Lee et al., 2024). These sources naturally contain long-horizon activities in which multiple response opportunities arise as the task progresses. Starting from the original task goal, step annotations, and temporal event labels, we convert each example into an interaction stream with one initial instruction and multiple response slots. Specifically, we rewrite the task topic or goal into a natural user instruction, synthesize it into speech using text-to-speech (Hu et al., 2026), and prepend the synthesized instruction to the original audio-visual stream. We then map step-level guidance targets or error events to temporally grounded response slots, each with its own answer time and target response. This procedure preserves the original video evidence while turning offline task annotations into an end-to-end audio-visual interaction setting, where the model receives the instruction through audio and must decide when to respond as new evidence appears. Benchmark examples are shown in Fig. 1 (1Q1A) and Fig. 2 (1QnA).

3.3 Evaluation Metrics

Continuous real-time human–AI interaction shifts evaluation from static correctness to dynamic interaction management. Traditional metrics are insufficient for online settings, particularly for handling full-duplex interruptions and nested context resumption. We therefore build our scoring framework upon the interaction slot formulation, anchoring evaluation to the triggers, response windows, and target answers introduced in Sec. 3.3.1 to jointly measure response timeliness, content quality, and conversational continuity.

3.3.1 Slot Construction and Chunk Matching

Continuous streams do not provide explicit turn boundaries, so we discretize evaluation into interaction slots:

$$\text{slot} = [t_{\text{start}}, t_a, t_{\text{end}}), \tag{1}$$

where t_{start} is the onset of observation, t_a is the earliest moment for a valid core response, and t_{end} is the window's close. Fig. 3 illustrates how slots are constructed across representative interaction types defined in Sec. 3.1.

We establish real-time and proactive interactions as the foundational structure: t_{start} aligns with the user query, t_a is the time of the visual event that enables a valid answer, and t_{end} is bounded by the subsequent query. For nested interactions, the outer slot keeps this definition, while the inserted query opens an inner slot that ends at $t_a(\text{outer})$, when the visual event makes the outer proactive response timely again and evaluation switches back to the outer slot. For 1QnA, which handles sequential responses to a single instruction, the first step follows the foundational structure. In subsequent steps, each visual event triggers the next slot, whose t_{start} and t_a align (labeled as t_{start}), and the next slot's t_{start} serves as current slot's t_{end}. Within these settings, a new user query or visual event of another

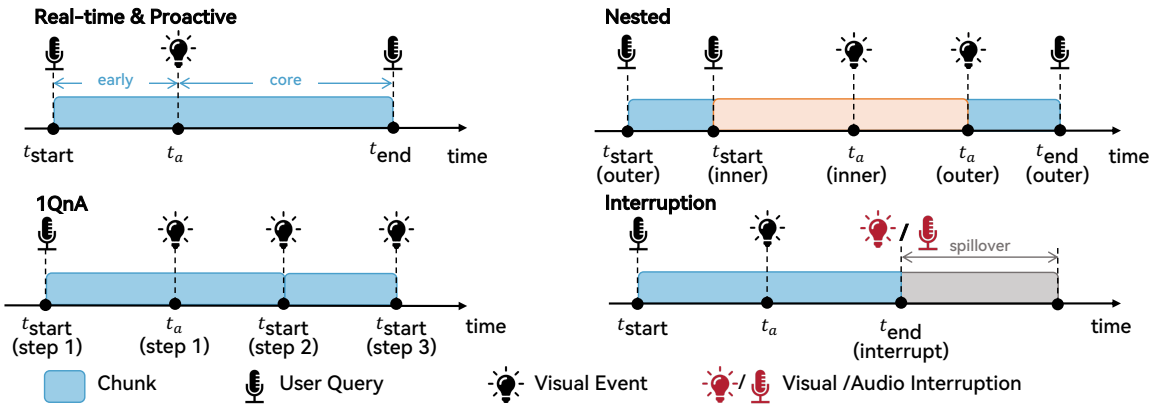


Figure 3: Interaction slot construction for real-time, proactive, nested, 1QnA, and interruption settings. Generated chunks are assigned to temporal slots and split into early and core segments around the valid-answer time for interaction-aware evaluation.

slot (which defines t_{end}) may arrive before the current answer is completed. We refer to this as an interruption, where the current slot is termed the interrupted slot, completing its response is not required, and any output after t_{end} is considered spillover. In practice, we annotate an interruption when the interval $[t_a, t_{\text{end}})$ is shorter than the TTS-estimated duration of the ground-truth answer.

Building on these definitions, a model-generated text chunk is assigned to a slot if its start time falls within $[t_{\text{start}}, t_{\text{end}})$. In cases of overlap, such as nested resumptions, the chunk is mapped to the slot with the latest t_{start} time, prioritizing the most recent context. Chunks straddling the t_a boundary are split at the word level into an early segment (before t_a) and a core segment (from t_a onward). Unassigned chunks are recorded as unmatched outputs and penalized during metric computation.

3.3.2 Interaction-Aware Scoring

For each slot, we compute a unified set of stage-specific scores to derive soft true positives (TP) and discrete penalties (FP, FN), integrating interaction management into a generalized framework.

Stage-Specific Scoring. We evaluate intra-slot outputs across an early stage ($t < t_a$) and a core stage ($t \geq t_a$), both incorporating a time-decay mechanism to reward promptness. **1)** The early stage evaluates tentative acknowledgments or feedback, where valid interactions are rewarded based on onset timing, while early hallucinations yield zero. **2)** The core stage assesses the correctness and coverage of the ground-truth answer, penalized by its latency relative to t_a. The total validity of an interaction is a soft true positive (TP), computed as the clamped sum of both stage scores. Full scoring definitions are provided in Appendix A.2.

Classification and Global Metric. Interaction failures are captured via discrete penalties. A false negative (FN) is assigned when a non-interruption slot lacks a core answer. A false positive (FP) aggregates four unwarranted behaviors: **1)** unmatched chunks, **2)** early hallucinations, **3)** low-quality responses, and **4)** spill, where output exceeds the boundary t_{end} to disrupt conversational continuity. Across all slots, we define the Interaction-Aware Quality-Timeliness F1 (IA-QTF1) as:

$$\text{IA-QTF1} = \frac{2 \cdot \sum TP}{2 \cdot \sum TP + \sum FP + \sum FN}. \quad (2)$$

By using soft TP values to account for response timing while penalizing flow-breaking behaviors like spill, IA-QTF1 provides a comprehensive assessment of a model's ability to manage dynamic multimodal dialogue.

3.3.3 Extended Metrics

To further assess specific interaction capabilities in greater detail, we define targeted metrics for interruption handling and nested context management.

Interruption Diagnostic Suite (IDS). Interrupted slots include both user-initiated interruptions, where the original answer is often no longer needed, and event-triggered shifts, where partial answers to the preempted query may still be useful. Because Global IA-QTF1 treats all interruptions as boundary-control cases and does not reward incomplete answers to the preempted query, the metric does not distinguish between silence, useful partial responses, and post-interruption spillover. IDS addresses this gap with three complementary diagnostics: **No-Output Rate (NOR)**, the proportion of interrupted slots with no model output for the preempted query; **Partial Answer Quality (PAQ)**, an

Table 3: IA-QTF1 across interaction settings. The 1Q1A columns use mutually exclusive real-time, proactive, and nested response slots; global scores are recomputed from aggregated TP/FP/FN.

Model	1Q1A				1QnA	All Global
	Real-time	Proactive	Nested	1Q1A Global		
AURA	0.376	0.549	0.596	**0.467**	**0.052**	0.363
Gemini 2.5 Flash Live	**0.553**	0.121	0.398	0.428	0.028	0.344
MiniCPM-o 4.5	0.337	**0.607**	**0.599**	0.456	0.015	**0.368**
Qwen3.5-Omni Flash Realtime	0.524	0.108	0.379	0.401	0.023	0.323

LLM-judged usefulness score for already-spoken content without incompleteness penalties; and **Conditional Spill Metrics (CSM)**, spill rate and average spill duration computed only over interrupted slots with output.

Nested Chain Completion Score. To evaluate state management during inserted queries, we further define the Nested Chain Completion Score (NCCS) as the geometric mean of correctness across the outer–inner query pair:

$$\text{NCCS} = \sqrt{\text{Score}_{\text{outer}} \times \text{Score}_{\text{inner}}}\,. \tag{3}$$

Here, $\text{Score}_{\text{outer}}$ and $\text{Score}_{\text{inner}}$ are outer/inner core-stage scores. NCCS requires answering the inner query and then resuming the outer query, measuring context-switching and resumption fidelity.

4 Experiments

We evaluate four representative omnimodal real-time models: AURA (Lu et al., 2026a), Gemini 2.5 Flash Live (Comanici et al., 2025), MiniCPM-o 4.5 (Cui et al., 2026), and Qwen3.5-Omni Flash Realtime (Team, 2026). All models are tested using their original real-time inference pipelines and native audio-visual streams, requiring them to jointly handle spoken user intents, visual evidence, and response timing. Since the answers are open-ended, we use GPT-4o (Hurst et al., 2024) as an external judge to compare model responses against ground-truth annotations, thereby reducing evaluator bias from the tested models. The judge protocol is detailed in Appendix A.4.

4.1 Inference Protocol

Although OmniInteract is distributed as offline audio-visual recordings for reproducible evaluation, all models are evaluated under an online streaming protocol. During inference, each recording is replayed chronologically to the model through its native real-time interface, so that frames and audio are exposed only according to their original timestamps. The model can therefore condition on past and current inputs, but cannot access future video frames, future audio, or ground-truth slot boundaries. We timestamp model outputs during replay and align the generated chunks with interaction slots after inference using the procedure in Sec. 3.3.1. This protocol simulates real online interaction while keeping the benchmark deterministic and comparable across models.

4.2 1Q1A Interaction

The 1Q1A split evaluates localized response opportunities, including explicit user queries, proactive triggers, and nested queries. Tab. 3 reports IA-QTF1 for each category and the global score.

For explicit real-time queries, Gemini obtains the best score (0.553), followed by Qwen3.5-Omni (0.524), showing stronger performance when the user intent is directly stated. In contrast, proactive interaction favors MiniCPM-o (0.607) and AURA (0.549), suggesting better monitoring after an earlier query whose answer becomes available only later. On nested slots, MiniCPM-o and AURA again perform best, indicating stronger local handling of context shifts. Under the global 1Q1A metric, which aggregates TP/FP/FN across all slots, AURA achieves the highest IA-QTF1 (0.467), slightly ahead of MiniCPM-o (0.456).

Nested IA-QTF1 measures local validity of inner and outer answers, but does not fully capture whether the model resumes the suspended outer query after the inserted query. We therefore report NCCS in Tab. 4. MiniCPM-o achieves the best NCCS of 0.284, followed by AURA at 0.270. Although Gemini and Qwen3.5-Omni answer many inner queries correctly, they fail to resume the outer query in 119 and 116 of 120 cases, respectively, indicating that current models often treat nested queries as permanent context switches rather than temporary interruptions requiring resumption.

4.3 1QnA Interaction

The 1QnA split evaluates continuous task monitoring, where a single instruction may require multiple temporally grounded responses. As shown in

Table 4: Nested interaction results over 120 nested pairs. NCCS measures chain-level completion, while Inner and Outer IA-QTF1 report local slot quality.

Model	NCCS	Inner IA-QTF1	Outer IA-QTF1	Missed Outer
AURA	0.270	0.595	0.599	54 / 120
Gemini 2.5 Flash Live	0.001	0.595	0.165	119 / 120
MiniCPM-o 4.5	**0.284**	0.587	**0.612**	55 / 120
Qwen3.5-Omni Flash Realtime	0.012	**0.702**	0.092	116 / 120

Table 5: Interruption Diagnostic. NOR: No-Output Rate; PAQ: Partial Answer Quality; CSM-SR: Conditional Spill Rate; CSM-AS: Conditional Average Spill.

Model	NOR	PAQ ↑	CSM-SR ↓	CSM-AS (s) ↓
AURA	79.17%	0.293	60.00%	1.879
Gemini	85.94%	0.370	**40.74%**	**0.312**
MiniCPM-o	53.65%	**0.571**	83.15%	10.067
Qwen-Omni	71.35%	0.361	41.82%	0.613

Table 6: Full-duplex capability degradation. We compare the mathematical reasoning quality of MiniCPM-o 4.5 in offline and online (full-duplex) settings.

Metric	Offline	Online	Δ Drop
Pure Quality Score	0.6833	0.3475	-0.3358

Tab. 3, all models perform substantially worse on 1QnA than on 1Q1A. AURA obtains the highest IA-QTF1 score of 0.052, but the absolute score remains low. This suggests that long-horizon interaction remains difficult, as models often miss intermediate response opportunities or respond at inappropriate times, even when they can handle isolated 1Q1A cases.

When aggregating both splits, MiniCPM-o obtains the highest overall Global IA-QTF1 score of 0.368, followed by AURA at 0.363. The small gap between the best models, together with the uniformly low 1QnA scores, suggests that current systems have not yet achieved robust general-purpose streaming interaction behavior across localized and long-horizon settings (detailed breakdown in Appendix A.3).

4.4 More Interruption Analyses

We use the Interruption Diagnostic Suite (IDS) defined in Sec. 3.3.3 to further separate no output for the preempted query from failed stopping behavior and to measure conditional spill severity. Tab. 5 shows that Gemini avoids spillover mostly through conservative silence, with the highest NOR (85.94%), modest PAQ (0.370), and the best CSM (40.74%, 0.312 s). MiniCPM-o shows the opposite pattern: it responds more often, with a lower NOR of 53.65% and the best PAQ of 0.571, but spills severely when it responds, with CSM of 83.15% and 10.067 s. Qwen3.5-Omni is more balanced, with NOR of 71.35% and relatively low CSM of 41.82% and 0.613 s, while AURA combines high silence (NOR 79.17%) with modest PAQ (0.293) and elevated spillover (CSM 60.00%, 1.879 s).

4.5 Full-duplex Capability Degradation

Finally, we examine whether offline capability transfers to online full-duplex-oriented interaction. We focus on MiniCPM-o 4.5, which is, to the best of our knowledge, the only open-source model that currently supports full-duplex real-time interaction. For offline inference, the entire question video is provided to MiniCPM-o at once, and the model answers after observing the full input. We compare its mathematical reasoning performance under offline inference and online full-duplex streaming interaction. To isolate answer correctness, we report the pure quality score (by GPT-4o), which excludes time decay and FP/FN penalties. As shown in Tab. 6, MiniCPM-o drops from 0.6833 offline to 0.3475 online, an absolute decrease of 0.3358. This suggests that continuous listening, visual processing, and concurrent response generation can substantially degrade reasoning quality. This result reinforces the need to evaluate omnimodal models in native streaming interaction, rather than relying solely on offline multimodal reasoning scores, highlighting the value of OmniInteract as a benchmark.

5 Conclusion

We introduced OmniInteract, a benchmark for evaluating omnimodal LLMs in native online streaming audio-visual interaction. Unlike offline or pre-segmented QA benchmarks, OmniInteract preserves spoken queries, visual events, ambient sounds, and response timing, enabling joint evaluation of answer quality, timeliness, interruption handling, and context resumption. Experiments show that current models struggle with robust real-time interaction, especially in long-horizon 1QnA monitoring and nested query resumption. These results highlight the gap between offline multimodal understanding and reliable full-duplex-oriented interaction, providing a foundation for future research on more natural human–AI communication.

Limitations

OmniInteract has several limitations that point to future work. First, we evaluate four representative models, but the landscape of omnimodal systems is evolving rapidly. Second, the online capability degradation analysis is limited to MiniCPM-o on mathematical reasoning tasks. Third, the 1QnA split uses TTS-synthesized speech for initial instructions, while 1Q1A queries are naturally recorded, which may introduce variation in speech recognition difficulty. Finally, the benchmark currently covers Chinese daily-life interactions and English mathematical reasoning, and broader language and domain coverage remains future work.

Ethical Considerations

OmniInteract is a research benchmark for evaluating real-time omnimodal interaction capabilities. It does not collect or release unauthorized personal user data; all self-recorded videos were created by the authors with informed consent from individuals who appear in them, and the 1QnA split builds on publicly available datasets under their original licenses. While real-time omnimodal assistants may support accessibility, education, and hands-free guidance, always-on multimodal systems also raise privacy and surveillance concerns that require careful deployment safeguards.

References

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, and 1 others. 2025. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*.

Apratim Bhattacharyya, Bicheng Xu, Sanjay Haresh, Reza Pourreza, Litian Liu, Sunny Panchal, Leonid Sigal, and Roland Memisevic. 2026. Can multi-modal llms provide live step-by-step task guidance? *Advances in Neural Information Processing Systems*, 38:22377–22410.

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

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. 2025. Livecc: Learning video llm with streaming speech transcription at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29083–29095.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, and 1 others. 2026. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, and 1 others. 2026. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*.

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

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, and 1 others. 2026. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shih-Po Lee, Zijia Lu, Zekun Zhang, Minh Hoai, and Ehsan Elhamifar. 2024. Error detection in egocentric procedural task videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18655–18666.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Guan-Ting Lin, Chen Chen, Zhehuai Chen, and Hung-yi Lee. 2026a. Full-duplex-bench-v3: Benchmarking tool use for full-duplex voice agents under real-world disfluency. *arXiv preprint arXiv:2604.04847*.

Guan-Ting Lin, Shih-Yun Shan Kuan, Jiatong Shi, Kai-Wei Chang, Siddhant Arora, Shinji Watanabe, and Hung-yi Lee. 2025a. Full-duplex-bench-v2: A multi-turn evaluation framework for duplex dialogue systems with an automated examiner. *arXiv preprint arXiv:2510.07838*.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H Liu, and Hung-yi Lee. 2025b. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. *arXiv preprint arXiv:2503.04721*.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. 2026b. 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)*, pages 12147–12151. IEEE.

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

Zikang Liu, Longteng Guo, Handong Li, Ru Zhen, Xingjian He, Ruyi Ji, Xiaoming Ren, Yanhao Zhang, Haonan Lu, and Jing Liu. 2026. Thinking in streaming video. *arXiv preprint arXiv:2603.12938*.

Xudong Lu, Yang Bo, Jinpeng Chen, Shuhan Li, Xintong Guo, Huankang Guan, Fang Liu, Dunyuan Xu, Peiwen Sun, Heyang Sun, and 1 others. 2026a. Aura: Always-on understanding and real-time assistance via video streams. *arXiv preprint arXiv:2604.04184*.

Xudong Lu, Huankang Guan, Yang Bo, Jinpeng Chen, Xintong Guo, Shuhan Li, Fang Liu, Peiwen Sun, Xueying Li, Wei Zhang, and 1 others. 2026b. Phostream: Benchmarking real-world streaming for omnimodal assistants in mobile scenarios. *arXiv preprint arXiv:2601.22575*.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, and 1 others. 2025. Ovo-bench: How far is your video-llms from real-world online video understanding? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18902–18913.

Rohith Peddi, Shivvrat Arya, Bharath Challa, Likhitha Pallapothula, Akshay Vyas, Bhavya Gouripeddi, Qifan Zhang, Jikai Wang, Vasundhara Komaragiri, Eric Ragan, and 1 others. 2024. Captaincook4d: A dataset for understanding errors in procedural activities. *Advances in Neural Information Processing Systems*, 37:135626–135679.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. 2025. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24045–24055.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. 2026. A simple baseline for streaming video understanding. *arXiv preprint arXiv:2604.02317*.

Meituan LongCat Team, Bin Xiao, Chao Wang, Chengjiang Li, Chi Zhang, Chong Peng, Hang Yu, Hao Yang, Haonan Yan, Haoze Sun, and 1 others. 2026. Longcat-next: Lexicalizing modalities as discrete tokens. *arXiv preprint arXiv:2603.27538*.

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

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. 2026. Streambridge: Turning your offline video large language model into a proactive streaming assistant. *Advances in Neural Information Processing Systems*, 38:132332–132359.

Yueqian Wang, Songxiang Liu, Disong Wang, Nuo Xu, Guanglu Wan, Huishuai Zhang, and Dongyan Zhao. 2025a. Mmduet2: Enhancing proactive interaction of video mllms with multi-turn reinforcement learning. *arXiv preprint arXiv:2512.06810*.

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

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. 2025c. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18925–18935.

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

Jiaer Xia, Peixian Chen, Mengdan Zhang, Xing Sun, and Kaiyang Zhou. 2025. Streaming video instruction tuning. *arXiv preprint arXiv:2512.21334*.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. 2025. Streamingvlm: Real-time understanding for infinite video streams. *arXiv preprint arXiv:2510.09608*.

Haolin Yang, Feilong Tang, Lingxiao Zhao, Xinlin Zhuang, Yifan Lu, Xiang An, Ming Hu, Xiaofeng Zhang, Abdalla Swikir, Junjun He, and 1 others. 2025. Streamagent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*.

Yiqun Yao, Xiang Li, Xin Jiang, Xuezhi Fang, Naitong Yu, Wenjia Ma, Aixin Sun, and Yequan Wang. 2025. Flm-audio: Natural monologues improves native full-duplex chatbots via dual training. *arXiv preprint arXiv:2509.02521*.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, and 1 others. 2026. Streamforest: Efficient online video understanding with persistent event memory. *Advances in Neural Information Processing Systems*, 38:75804–75835.

Yulin Zhang, Cheng Shi, Yang Wang, and Sibei Yang. 2025. Eyes wide open: Ego proactive video-llm for streaming video. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, and 1 others. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489.

A Appendix

A.1 Data Licenses and Annotation Details

Tab. A.1 summarizes the licenses and access terms for the external data sources and data-generation tools. For human annotation, annotators were compensated at a rate of US$20 per hour.

A.2 Detailed Scoring Definitions

Early Stage Score ($\text{Score}_{\text{ack}}$). Within the early segment $[t_{\text{start}}, t_a)$, model outputs are evaluated for appropriate interaction behavior. Valid acknowledgments (*e.g.*, confirmations, brief feedback, or wait signals) are rewarded with a score that decays with onset latency relative to the early window length, scaled by a cap factor α to ensure that acknowledgments contribute less than core answers. If the model instead produces an early hallucination (*i.e.*, a substantive answer before sufficient evidence has emerged at t_a), the acknowledgment score is set to zero and a false positive is recorded.

Core Stage Score ($\text{Score}_{\text{core}}$). Within the core segment $[t_a, t_{\text{end}})$, the score combines a semantic quality factor and a timeliness factor:

$$\text{Score}_{\text{core}} = S_{\text{core}} \times T_{\text{core}}, \tag{4}$$

where $S_{\text{core}} \in [0, 1]$ is the semantic quality score assigned by the LLM judge (Sec. A.4), assessing correctness and coverage against the ground-truth answer. $T_{\text{core}} \in [0, 1]$ is a timeliness factor that decays linearly from 1 to 0 as the semantic anchor (*i.e.*, the earliest chunk containing the key answer content, as identified by the judge) shifts from t_a toward t_{end}:

$$T_{\text{core}} = \max\left(0,\ 1 - \frac{t_{\text{anchor}} - t_a}{t_{\text{end}} - t_a}\right). \tag{5}$$

Soft True Positive. The per-slot soft TP is defined as:

$$TP_n = \min(1,\ \text{Score}_{\text{ack}} + \text{Score}_{\text{core}})\ . \tag{6}$$

The clamping ensures the combined score does not exceed 1.

False Positive Categories. Each slot may incur FP counts from four sources: (1) unmatched chunks not assigned to any slot, (2) early hallucinations in the $[t_{\text{start}}, t_a)$ segment, (3) core responses with quality below a minimum threshold, and (4) spillover output beyond the slot boundary t_{end}.

False Negative. A non-interrupted slot is assigned $FN = 1$ when $\text{Score}_{\text{core}} \leq 0$, *i.e.*, the model fails to produce any valid core answer within the response window. Acknowledgments alone do not satisfy the completion requirement. Interrupted slots do not incur FN, since the interaction was preempted before the model was expected to complete its answer.

More Interruption Diagnostics. For interrupted slots, the global IA-QTF1 score only checks boundary control and does not require completing the original answer. We therefore report separate diagnostics. **No-Output Rate (NOR)** is the fraction of interrupted slots with no model output. For interrupted slots with output, **Partial Answer Quality (PAQ)** is an LLM-judged score in $[0, 1]$ measuring whether the already spoken partial response is relevant, correct, and useful; incompleteness alone is not penalized. **Conditional Spill Metrics (CSM)** measure spill rate and average spill duration only over interrupted slots with output.

A.3 Detailed TP/FP/FN Breakdown

Tab. A.2 reports the per-category TP, FP, and FN values underlying the IA-QTF1 scores in Tab. 3. The 1Q1A categories use mutually exclusive response slots; the global score aggregates all 1,430 slots and includes unmatched-chunk FP that are not attributed to any individual category.

Table A.1: Licenses and access terms for external data sources and data-generation tools used in OmniInteract.

Source	Use in OmniInteract	License / access terms	Notes
Qualcomm Interactive Cooking Dataset (Bhattacharyya et al., 2026)	1QnA live step-by-step task guidance instances	Data License Agreement – Research Use	The dataset provides manually annotated instructions, feedback messages, and timestamps; its videos are from CaptainCook4D.
CaptainCook4D (Peddi et al., 2024)	Procedural activity videos and annotations used directly and through the Qualcomm Interactive Cooking Dataset	Apache License 2.0 for the dataset; MIT License for the public annotation repository	We follow the license terms of the specific files used.
EgoPER (Lee et al., 2024)	1QnA egocentric procedural error-detection instances	Available by request under the original dataset owners' access terms	The official project page requires users to request dataset access from the authors and provide institutional and research-purpose information.
Qwen3-TTS (Hu et al., 2026)	Text-to-speech synthesis for initial 1QnA spoken instructions	Apache License 2.0	Used only to synthesize the initial instruction audio prepended to the original audio-visual streams.

Table A.2: Detailed TP/FP/FN breakdown per interaction category. 1Q1A categories are mutually exclusive. "All Global" aggregates all 1,430 slots; its FP includes unmatched chunks not attributed to individual categories.

Model	Category	TP	FP	FN	Precision	Recall	IA-QTF1
AURA	Real-time	162.42	293	245	0.357	0.399	0.376
	Proactive	90.59	95	54	0.488	0.627	0.549
	Nested	116.77	73	85	0.615	0.579	0.596
	1QnA	12.52	159	294	0.073	0.041	0.052
	All Global	**382.30**	**666**	**678**	**0.365**	**0.361**	**0.363**
Gemini 2.5 Flash Live	Real-time	211.32	163	179	0.565	0.541	0.553
	Proactive	16.77	73	170	0.187	0.090	0.121
	Nested	69.06	50	159	0.580	0.303	0.398
	1QnA	4.97	37	314	0.118	0.016	0.028
	All Global	**302.12**	**328**	**822**	**0.480**	**0.269**	**0.344**
MiniCPM-o 4.5	Real-time	150.10	326	264	0.315	0.363	0.337
	Proactive	97.27	62	64	0.611	0.603	0.607
	Nested	119.65	72	88	0.624	0.576	0.599
	1QnA	2.92	70	316	0.040	0.009	0.015
	All Global	**369.94**	**539**	**732**	**0.407**	**0.336**	**0.368**
Qwen3.5-Omni Flash Realtime	Real-time	216.12	218	174	0.498	0.554	0.524
	Proactive	17.23	113	171	0.132	0.092	0.108
	Nested	76.88	112	140	0.407	0.354	0.379
	1QnA	4.41	68	314	0.061	0.014	0.023
	All Global	**314.64**	**522**	**799**	**0.376**	**0.283**	**0.323**

A.4 LLM Judge Evaluation Protocol

All open-ended answer assessments use GPT-4o (Hurst et al., 2024) as an external judge to avoid evaluator bias from the tested models. Core-stage assessment receives: (1) the ground-truth target answer, (2) the concatenated model-generated chunks within the core segment, and (3) a structured instruction asking it to rate semantic correctness and coverage on a continuous scale of $[0, 1]$. The judge also identifies the semantic anchor (*i.e.*, the earliest chunk that contains the key answer content) used to compute the timeliness factor T_{core}. Early-stage assessment is performed separately on chunks before t_a, where the judge classifies outputs as either valid acknowledgments (brief interaction feedback) or early hallucinations (premature substantive content). For 1QnA slots, the judge additionally checks whether the model reveals information about future steps before they become relevant, flagging such outputs as spoilers. For interrupted slots with output, the judge scores the Partial Answer Quality (PAQ) of the already spoken content without penalizing incompleteness.

The judge uses separate prompts for early-stage, core-stage, and interruption-diagnostic scoring. Listing A.1 classifies outputs in $[t_{\text{start}}, t_a)$ as neutral acknowledgments or early hallucinations. Listing A.2 scores the usefulness of already spoken partial outputs without penalizing incompleteness. Listing A.3 scores core-answer quality in $[t_a, t_{\text{end}})$ and extracts the trigger phrase for semantic anchor identification. The prompts shown below are English translations of the original Chinese prompts used in evaluation.

Listing A.1: Early-stage judge prompt template.

```
[System]
You are a streaming voice assistant evaluation judge. Judge only based on the given text. Output must be parseable JSON with no
other text.

[User]
Determine whether the early output between start and t_a is an early hallucination.

[scene_type] {scene_type}
[slot] slot_id={slot_id},
  turn_index={turn_index},
  step_index={step_index},
  boundary_type={boundary_type},
  is_interrupted={is_interrupted}
[question] {question}
[current_gt_answer] {gt_answer}
[full_chunk_context] {full_context}
[early_actual_text] {actual_text}

Rules:
1. Greetings, confirmations, waiting, brief observations, and follow-up phrases -> Neutral.
2. If the model starts substantively answering, guessing unseen info, revealing future steps, or making definitive factual claims
-> FP.
3. For 1QnA first step, reciting the full procedure before acting -> FP.
4. score is interaction quality 0-1 when Neutral; 0 when hallucination.

Output JSON:
{"flag":"Neutral|FP_Hallucination",
 "score":float 0-1,
 "rationale":"one sentence"}
```

Listing A.2: Interrupted partial-answer quality prompt template.

```
[System]
You are a strict evaluator for interrupted voice-assistant answers. Judge only from the provided text. Return valid JSON only,
with no extra text.

[User]
Evaluate the quality of the assistant output that was already spoken before or around an interruption.

[Task]
The assistant was answering, but the interaction was interrupted before completion. The assistant was not required to complete the
 full original answer. Score whether the content already spoken is relevant, correct, and useful for the current ground-truth
answer.
```

```
[Question] {question}
[Ground Truth Answer] {gt_answer}
[Assistant Output Already Spoken]
{actual_text}

Scoring Rules:
1. Score from 0 to 1.
2. Do not penalize incompleteness: a partial answer can receive a high score if the spoken part is correct and useful.
3. Score high when the spoken content overlaps with, paraphrases, or conveys useful parts of the ground truth.
4. Score low for acknowledgments or prefaces without substantive answer content.
5. Score low for wrong-question, irrelevant, or generic-filler output.
6. hallucination=true if the output contains clear incorrect facts, wrong target content, or unsupported content.
7. Ignore overflow duration when scoring quality; spill is measured separately.

Output JSON:
{"score":float 0-1,
 "hallucination":true|false,
 "rationale":"one sentence"}
```

Listing A.3: Core-stage judge prompt template.

```
[System]
You are a strict streaming voice assistant core-answer evaluation judge. Judge only based on the given text and reference answer.
Output must be parseable JSON with no other text.

[User]
Score the core output after t_a.

[scene_type] {scene_type}
[slot] slot_id={slot_id},
  turn_index={turn_index},
  step_index={step_index},
  boundary_type={boundary_type},
  is_interrupted={is_interrupted}
[question] {question}
[current_gt_answer] {gt_answer}
[future_gt_answers_or_steps]
  {future_answers}
[full_chunk_context] {full_context}
[core_actual_text] {actual_text}

Rules:
1. score 0-1: correctness and coverage of core_actual_text vs gt_answer.
2. Off-topic, factual errors, or missing key answer -> low score.
3. 1QnA: reward only current-step info; penalize spoiling future steps or skipping the current step.
4. If score > 0, extract the earliest contiguous substring from core_actual_text that establishes the answer as trigger_phrase.
5. trigger_phrase must be a verbatim substring; empty if score == 0.

Output JSON:
{"score":float 0-1,
 "trigger_phrase":"substring or empty",
 "spoiler":true|false,
 "rationale":"one sentence"}
```

A.5 Case Study

We provide qualitative examples in Figs. A.1–A.5 to illustrate the behaviors behind the aggregate results in Sec. 4. Each example shows sampled video frames, the annotated interaction slot, the reference answer, and model outputs aligned to early/core segments. These cases make two patterns visible: models often possess the local perceptual ability needed to answer a frame-grounded question, but they frequently fail when the interaction requires deciding when to speak, when to wait, when to stop, or when to resume a suspended goal.

Annotation convention. The TP/FP/FN tags in the case figures denote slot- or stage-level outcomes, not independent per-chunk judgments. When several chunks belong to the same evaluated stage, the tag is placed on the last chunk to summarize the concatenated response judged for that stage. A "spill" tag indicates output beyond a hard interruption or slot boundary and may add an FP; 1QnA slots use soft boundaries, so slight carry-over between adjacent steps is tolerated and is not by itself counted as a spill FP. PAQ denotes the Partial Answer Quality score for interrupted slots, measuring the usefulness of the already spoken partial response without requiring completion.

Real-time visual question answering. In Fig. A.1, the user asks for the refrigerator's energy-efficiency level in the slot $[04{:}06, 04{:}11, 05{:}01]$. The visual evidence is localized: the label becomes readable around the valid-answer time, and the correct response is that the refrigerator is level 1 energy efficient. Gemini and Qwen3.5-Omni produce acceptable core answers, with TP scores of 0.7935 and 0.7123, respectively. In contrast, AURA and MiniCPM-o both answer that the refrigerator is level 2 energy efficient, yielding TP scores of 0.0000 with FP/FN penalties. This case supports the observation in Sec. 4 that explicit real-time queries are relatively easier than stateful interactions, but also shows that localized perception can still fail when the model misreads the fine-grained visual attribute.

Proactive response timing. Fig. A.2 shows a proactive book-title query: the user asks the assistant to report the title when a book appears, and the correct title is *The Stranger*. AURA waits with an acknowledgment and answers after the book becomes visible, achieving a TP score of 0.9343. MiniCPM-o behaves similarly and obtains a TP score of 0.8664. In contrast, Gemini responds in the early stage that no book is visible and asks the user to try again, while Qwen3.5-Omni prematurely guesses *The Little Prince*. Both are counted as early hallucinations and receive FP/FN penalties. The example explains why proactive IA-QTF1 favors MiniCPM-o and AURA in Tab. 3: success depends less on recognizing the final object alone and more on suppressing premature answers until the trigger is actually supported by the stream.

Nested context switching and resumption. The nested case in Fig. A.3 combines an outer monitoring instruction (notify the user when a kettle appears) with an inserted inner query asking for the title of a visible book. MiniCPM-o answers the inner question immediately, then resumes the outer task when the kettle appears, yielding successful NCCS with a score of 0.7845. AURA also completes both parts with NCCS 0.7593, although it uses more descriptive wording. Gemini fails because it treats the inner query as if the outer kettle task were still the active question, producing an early response about not seeing a kettle instead of reading the book title. Qwen3.5-Omni answers the inner book-title question correctly, but never resumes the outer monitoring task, so NCCS is zero despite a valid inner answer. This qualitative pattern matches Tab. 4: many models can answer the inserted query locally, but maintaining a suspended outer intent and returning to it remains difficult.

Interruption control. Fig. A.4 isolates the full-duplex stopping problem. The model is asked to read eleven public-kitchen rules, but the slot is interrupted at $02{:}11$, so completion is not required; the key behavior is whether generation stops at the boundary, while the partial content quality indicates whether the model has provided useful information before or around the interruption. Gemini stops before the interruption and has no spill, but its output is mostly a preface rather than the requested rules, yielding a low PAQ score of 0.20. Qwen3.5-Omni and AURA both read useful rule content and continue only slightly beyond the boundary, with PAQ scores of 0.85 and 0.90 and spill durations of 0.43 s and 1.54 s, respectively. MiniCPM-o also provides substantive rule content (PAQ 0.80), but continues reading for about 23 s after interruption, crossing the boundary with a long answer. This case directly supports the interruption diagnostics in Tab. 5: no spill alone can mask a lack of useful partial content, while high partial quality must still be considered together with conditional spill behavior.

Long-horizon 1QnA monitoring. The 1QnA

example in Fig. A.5 asks the model to guide a breakfast-burrito procedure and report mistakes. The first valid instruction is to crack an egg into a microwave-safe bowl; later slots include detecting eggshells in the bowl, prompting the user to whisk the egg, and then microwaving while stirring. All models struggle across the first four response slots. Gemini gives an irrelevant dishwasher-related response and then misses later slots. Qwen3.5-Omni answers with a generic skillet-based recipe, revealing unsupported future steps instead of tracking the observed procedure. MiniCPM-o produces a long monologue that rolls multiple future actions into one response, causing spill and losing temporal alignment. AURA is the only model with a nonzero score in the shown slots, but its valid response is delayed and partial, and it still misses the error-correction and next-step guidance. This case illustrates why all models have very low 1QnA IA-QTF1 in Sec. 4: continuous task assistance requires a sequence of small, timely decisions, so one early over-generation or missed event can degrade multiple slots.

Overall, the cases show that OmniInteract penalizes failures that are central to real streaming assistance: guessing before the evidence appears, missing when to respond, forgetting a paused request, and continuing after interruption. They therefore provide qualitative support for the main experimental conclusions: explicit localized queries are comparatively tractable; proactive and nested interactions expose state-management weaknesses; interruption handling varies sharply across models; and long-horizon 1QnA remains the most challenging setting.

04:06 04:07 04:11 04:18 05:01

Question Slot: [04:06, 04:11, 05:01]
Question: 请你告诉我这个冰箱是几级能耗的。
Answer : 好的，这个冰箱是一级能耗的。

Qwen3.5-Omni Flash Realtime

Answer:
(04:11) 这台冰箱是1级能效。 **(core, TP)**
Slot F1: 1.000
Score summary: TP=0. 7123, FP=0, FN=0

Gemini 2.5 Flash Live

Answer:
(04:11) 好的, 这台海尔冰箱是1级能耗的。还有什么想知道的吗? **(core, TP)**
Slot F1: 1.000
Score summary: TP=0.7935, FP=0, FN=0

MiniCPM-o 4.5

Answer:
(04:10) 这是海尔冰箱，它是二级能效的。 **(early & core, FP)**
Slot F1: 0.000
Score summary: TP=0.0000, FP=2, FN=1

AURA

Answer:
(04:07) 冰箱门顶部的右上方清晰地印有“Haier”字样。 **(early & core, FP)**
(04:12) 没问题，我帮你盯着，冰箱的能耗等级一确定就告诉你。 **(core)**
(04:18) 这个冰箱的能效等级是二级。可以看到冰箱门上贴着蓝色的能效标识，上面有一个醒目的绿色“2”，旁边还有从绿色到红色的阶梯状能效图。 **(core, FP)**
Slot F1: 0.000
Score summary: TP=0.0000, FP=2, FN=1

Figure A.1: Real-time interaction case. The user asks for the energy-efficiency level of a Haier refrigerator after the label becomes visible.

02:25 02:51 02:52 02:55 03:47

Question Slot: [02:25, 02:51, 03:47]
Question: 当视频中出现一本书的时候请你告诉我这本书的书名?
Answer : 这本书叫做《局外人》。

Qwen3.5-Omni Flash Realtime

Answer:
(02:31) 书名是《小王子》。 **(early, FP)**
(02:32-03:47) ***Silent***
Slot F1: 0.000
Score summary: TP=0.0000, FP=1, FN=1

Gemini 2.5 Flash Live

Answer:
(02:32) 抱歉,视频中没有看到任何书。 你能再试一次,或者告诉我书的信息吗? **(early, FP)**
(02:33-03:47) ***Silent***
Slot F1: 0.000
Score summary: TP=0.0000, FP=1, FN=1

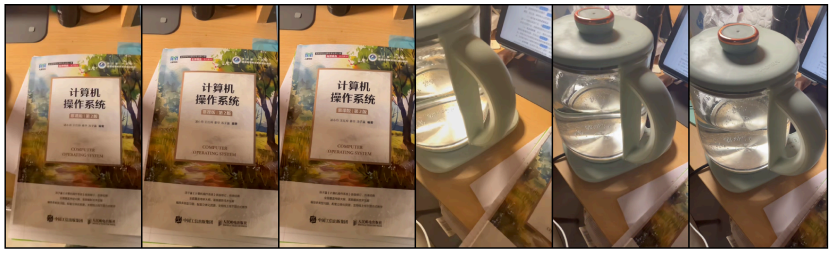
MiniCPM-o 4.5

Answer:
(02:30) 好的。 **(early, none)**
(02:55) 书的名字叫局外人。 **(core, TP)**
Slot F1: 1.000
Score summary: TP=0.8664, FP=0, FN=0

AURA

Answer:
(02:32) 没问题，等视频里一出现书我就告诉你书名。 **(early, none)**
(02:52) 镜头移动到桌面的另一侧，画面中出现了一本深蓝色的书，书的封面上清晰地印着“局外人”三个字。 **(core, TP)**
Slot F1: 1.000
Score summary: TP=0.9343, FP=0, FN=0

Figure A.2: Proactive interaction case. The model must wait until a book appears and then report its title.



Question Slot (Outer): [00:00, 00:55, 01:21]
Question (Outer): **当视频中出现一个烧水壶的时候请你告诉我。**
Answer (Outer): **烧水壶出现了。**
Question Slot (Inner): [00:08, 00:11, 00:55]
Question (Inner): **请你告诉我，面前这本书的书名是什么。**
Answer (Inner): **这本书书名叫做《计算机操作系统》。**

Qwen3.5-Omni Flash Realtime

Answer (Outer):
(00:05) 我现在还没看到烧水壶呢。 **(early, none)**
(00:55-01:21) ***Silent***
Slot F1: 0.155
Score summary: TP=0.0915, FP=0, FN=1

Answer (Inner):
(00:12) 这本书叫《计算机操作系统》，是第2版。 **(core, TP)**
Slot F1 : 1.000
Score summary: TP=0. 7895, FP=0, FN=0

NCCS: 0.0000
NCCS status: failed

Gemini 2.5 Flash Live

Answer (Outer):
(00:05) 好的, 我会帮你留意。视频中出现烧水 **(early, none)**
(00:55-01:21) ***Silent***
Slot F1: 0.225
Score summary: TP=0. 1452, FP=0, FN=1

Answer (Inner):
(00:10) 抱歉, 我目前还没有看到烧水壶。 你能再往周围看看吗? **(early, FP)**
Slot F1 : 0.000
Score summary: TP=0.0000, FP=2, FN=1

NCCS: 0.0000
NCCS status: failed

MiniCPM-o 4.5

Answer (Outer):
(00:04) 好的。 **(early, none)**
(00:56) 注意，视频中出现了一个烧水壶。 **(core, TP)**
Slot F1: 1.000
Score summary: TP=0.9176, FP=0, FN=0

Answer (Inner):
(00:11) 这本书的书名是计算机操作系统。 **(core, TP)**
Slot F1 : 1.000
Score summary: TP=0.8000, FP=0, FN=0

NCCS: 0.7845
NCCS status: success

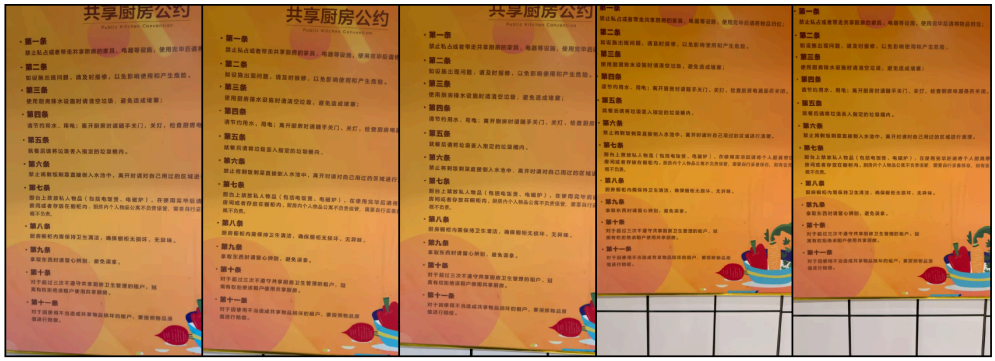
AURA

Answer (Outer):
(00:06) 好的，我会注意。 **(early, none)**
(00:57) 镜头移动至桌面左侧，展示了一个透明玻璃材质的烧水壶，壶身带有淡绿色的把手和壶盖。 **(core, TP)**
Slot F1: 1.000
Score summary: TP=0.9266, FP=0, FN=0

Answer (Inner):
(00:13) 收到，等我看到书名就立刻告诉你。 **(core, none)**
(00:17) 书本封面的中央清晰地印着黑色的书名“计算机操作系统”，下方还有一行较小的字体标注着“慕课版 第2版”，明确指出了这是一本关于计算机操作系统的教科书。 **(core, TP)**
Slot F1 : 1.000
Score summary: TP=0.7720, FP=0, FN=0

NCCS: 0.7593
NCCS status: success

Figure A.3: Nested interaction case. The model first monitors for a kettle, then answers an inserted book-title question, and finally should resume the outer monitoring task.



Question Slot: [01:11, 01:17, 02:11]
Question: **这个厨房有11条公约，你帮我念一遍**
Answer: **第一条：禁止私占或者带走共享厨房的家具、电器等设施，使用完毕后请将物品归位；第二条：如设施出现问题，请及时报修，以免影响使用和产生危险。第三条：使用厨房排水设施时请清空垃圾，避免造成堵塞；第四条：请节约用水、用电；离开厨房时请随手关门、关灯，检查厨房电器是否关闭。第五条：就餐后请将垃圾丢入指定的垃圾桶内。第六条：禁止将剩饭剩菜直接倒入水池中，离开时请对自己用过的区域进行清理。第七条：厨台上禁放私人物品（包括电饭煲、电磁炉），在使用完毕后请将个人厨具带回房间或者存放在橱柜内，厨房内个人物品公寓不负责保管，需要自行妥善保存，如有丢失，概不负责。第八条：厨房橱柜内需保持卫生清洁，确保橱柜无损坏、无异味。第九条：拿取东西时请留心辨别，避免误拿。第十条：对于超过三次不遵守共享厨房卫生管理的租户，公寓有权拒绝该租户使用共享厨房。第十一条：对于因使用不当造成共享物品损坏的租户，要按照物品原值进行赔偿。**

Qwen3.5-Omni Flash Realtime

Answer:
([01:18, 02:11]) 第一条，禁止私占或者带走共享厨房的家具、电器等设施，使用完毕后请将物品归位。第二条，如设施出现故障，请及时报修，以免影响使用和产生危险。第三条，使用厨房排水设施时请清空垃圾，避免造成堵塞。第四条，请节约用水、用电，离开厨房时请随手关门、关灯，检查厨房电器是否关闭。第五条，就餐后将残渣投入指定的垃圾桶内。第六条，禁止将剩饭剩菜直接倒入水池中，离开时请对自己用过的区域进行清洁。第七条，厨房内私放个人物品（包括电饭煲、微波炉），在使用完毕后请将物品存放在橱柜内，厨房内人物较多 **(PAQ=0.85)**
spill=True
spill_seconds=0.43

Gemini 2.5 Flash Live

Answer:
([01:19, 01:26]) 好的, 没问题。 这边的《共享厨房公约》共有11条, 我给您念一遍。您准备好了吗? **(PAQ=0.20)**
spill=False
spill_seconds=0

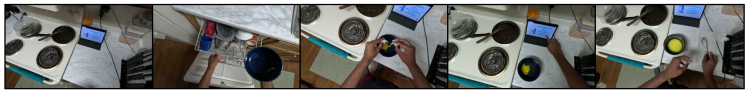
MiniCPM-o 4.5

Answer:
([01:16, 02:34]) 好的，我们一起来念一下第一条，禁止私占或者带走共享厨房的家具电器等设施。第二条如果设施出现问题请及时报修，以免影响使用和产生危险。第三条使用厨房排水设施时请清空垃圾，避免造成堵塞第四条节约用水用电离开厨房时请随手关门关灯检查电器是否漏电第五条就餐后将垃圾丢入指定的垃圾桶内第六条禁止将剩饭剩菜直接倒入水池中离开时请对自己所用过的区域进行清理第七条台面上禁止放私人物品包括电饭煲电水壶在使用完毕后请将个人物品原样带回家如果存放在橱柜内则需锁好柜门第八条厨房。柜内需保。持卫生清洁确保柜无坏损无异味第九条享受东西时请留心周围避免误拿第十条对于超过三次不遵守共享厨房卫生管理的租户冠以有效提示十一条对于因使用不当造成共享物品损坏的租户要按照物品价值进行赔偿。 **(PAQ=0.80)**
spill=True
spill_seconds=23

AURA

Answer:
([01:19, 01:21]) 好嘞，我给你念这十一条公约。
([01:24, 02:13]) 第一条，禁止私自占用或者带走共享厨房的家具、电器等设施，用完后请归还。第二条，如设施出现问题，请及时报修，以免影响使用和产生危险。第三条，使用厨房排水设施时请清空垃圾，避免造成堵塞。第四条，节约用水、用电；离开厨房时请随手关门、关灯，检查厨房电器是否关闭。第五条，就餐后请将垃圾丢入指定的垃圾箱内。第六条，禁止将剩饭剩菜直接倒入水池中，离开时请对自己 **(PAQ=0.90)**
spill=True
spill_seconds=1.54

Figure A.4: Interruption case. The user asks the model to read public-kitchen rules, but the answer window is truncated by an interruption.



00:00 00:08 00:45 01:08 01:56

Question: I'm planning to make some breakfast burritos. Could you guide me through the steps and let me know if I make any mistakes along the way? Let's start now.
Step1 Slot: [00:00, 00:08, 00:45]
Step1: Sure! Let's start by cracking an egg into a microwave-safe bowl.
Step2 Slot: : [00:45, 01:08]
Step2: Careful, some eggshells fell into the bowl. You'll want to fish those out.
Step3 Slot: [01:08, 01:56]
Step3: Now, go ahead and whisk the egg.

Qwen3.5-Omni Flash Realtime

Step1 Answer:
([00:09, 00:41]) Alright, let's get started. First thing you'll need is a large skillet or frying pan. Turn your stove on to medium heat and add a little oil or butter to the pan. While that's heating up, grab some eggs and crack them into a bowl. You can beat them with a fork if you want, but it's not necessary. Once your pan is hot, pour in the eggs. Let them cook for a minute or two until the edges start to set, then gently scramble them with a spatula. Keep stirring until they're cooked through but still soft. (core, FP)
Slot F1: 0.000
Score summary: TP=0.0000, FP=1, FN=1
Step2 Answer:
([00:45, 01:08]) *Silent*
Slot F1: 0.000
Score summary: TP=0.0000, FP=0, FN=1
Step3 Answer:
([01:08, 01:56]) *Silent*
Slot F1: 0.000
Score summary: TP=0.0000, FP=0, FN=1

Gemini 2.5 Flash Live

Step1 Answer:
([00:09, 00:16]) I can definitely help with that! Let's start by getting the ingredients you'll need. It looks like you're at the dishwasher, did you need a plate or bowl? (core, FP)
Slot F1: 0.000
Score summary: TP=0.0000, FP=1, FN=1
Step2 Answer:
([00:45, 01:08]) *Silent*
Slot F1: 0.000
Score summary: TP=0.0000, FP=0, FN=1
Step3 Answer:
([01:08, 01:56]) *Silent*
Slot F1: 0.000
Score summary: TP=0.0000, FP=0, FN=1

MiniCPM-o 4.5

Step1 Answer:
([00:07, 01:54]) Sure, I will be here for you. First of all, open the dishwasher and get a bowl out. Then head over to the stove area. It looks like you are about to put the bowl down there. But then you changed your mind and headed towards the sink instead. You opened the fridge door next. What's in that box? Oh, it seems to be some frozen food items. Place them on the countertop carefully. And then take out an egg from the carton. Now go back to the stove area again. Put the egg into the bowl and crack it open with your fingers. That sounds like a fresh egg though. Make sure not to let any shell pieces fall into the bowl as they can be hard to remove later. After cracking the egg, dispose of the shells properly. And grab a paper towel to clean up any messes around the countertop. Then return the box to the fridge. Don't forget to close the fridge door afterwards too. Next, look at the recipe instructions on your tablet. Are you going to whisk the eggs now? Yes, you took out a whisk and started to beat the egg. Keep whisking until the egg is well mixed. The sound of whisking is so satisfying. Do this step slowly but steadily. I hope you're enjoying making these bre (core, FP) (spill)
Slot F1: 0.000
Score summary: TP=0.0000, FP=1, FN=1
Step2 Answer:
([00:45, 01:08]) *Silent*
Slot F1: 0.000
Score summary: TP=0.0000, FP=0, FN=1
Step3 Answer:
([01:08, 01:56]) *Silent*
Slot F1: 0.000
Score summary: TP=0.0000, FP=0, FN=1

AURA

Step1 Answer:
([00:10, 00:17]) Got it, I'll guide you step by step and let you know if you make any mistakes. (core)
([00:22, 00:38]) I'm opening the refrigerator to see what's inside. The first thing I'm looking at is the top shelf, where there's a container of butter and a jar of what looks like mayonnaise or a similar dressing. (core)
([00:38, 00:50]) I've taken down a box of rice from the top shelf of the refrigerator. The box is white with black text, and I can see the words "Rice and Vegetables" on it. (core, FP) (spill)
Slot F1: 0.000
Score summary: TP=0.0000, FP=1, FN=1
Step2 Answer:
([00:50, 01:04]) I'm opening the box that I just took out. It's a cardboard box, and as I lift the lid, I can see there are plastic containers inside. (core)
([01:04, 01:19]) The person opens the cardboard box, revealing two layers of egg cartons. They take out a carton of eggs, which are white, and then close the box, leaving it on the counter. (core, FP) (spill)
Slot F1: 0.000
Score summary: TP=0.0000, FP=1, FN=1
Step3 Answer:
([01:19, 01:34]) The person picks up a white egg from their hand. They then move towards the stove, holding one egg in each hand, and place them into a pan that is already on one of the burners. (core)
([01:34, 01:43]) The person is holding a blue bowl in their left hand. Using their right hand, they break an egg into this bowl. (core)
([01:43, 01:59]) I've just cracked an egg into the blue bowl. As I do this, I can see the tablet on the counter next to me, which has what looks like a grocery list displayed on it. I'm thinking about what to cook now. (core, TP) (spill)
Slot F1: 1.000
Score summary: TP=0.3808, FP=0, FN=0

Figure A.5: 1QnA case. A breakfast-burrito instruction requires multiple temporally grounded responses across a continuous procedure.