OMNIDUET-BENCH: DO OMNI-MODAL MODELS ACT ON WHAT THEY SEE IN FULL-DUPLEX AUDIOVISUAL CONVERSATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-time omni-modal models can now generate speech while receiving streaming audio and video, yet no benchmark systematically evaluates whether they respond to the speech, facial expressions, and actions in their audiovisual input with appropriate timing and content: full-duplex benchmarks evaluate only spoken interaction, and video understanding benchmarks assess comprehension of video content rather than how a model, as a conversational participant, responds to the user on screen. We present OmniDuet-Bench, which evaluates full-duplex audiovisual conversation along three dimensions: Interaction Capacity, Visual Perception, and Cross-Modal Integration. The benchmark comprises 707 human-annotated task instances curated from natural two-person video calls and organized into six streaming tasks. Visual Perception and Cross-Modal Integration tasks share the same expression and action events and are complemented by matched audiovisual and audio-only conditions; together, these designs dissociate the perception of a visual cue from its use in conversational behavior. We evaluate seven real-time omni-modal models and find that no model leads on all three dimensions: models often fail to hold the floor and to take the turn promptly when the user yields it, and even after identifying an expression or action correctly, they often fail to adapt their responses to it. Data and evaluation code will be released.

1 INTRODUCTION

Video conversation unfolds concurrently through speech and visible behavior: participants listen to each other while attending to facial expressions and body movements. They use these verbal and nonverbal cues to regulate turn-taking and to adapt what they say and how they say it. Real-time omni-modal models such as Gemini Live, GPT Realtime, Qwen3-Omni, and MiniCPM-o 4.5 can now generate speech while receiving audiovisual input (Google Cloud, 2026; OpenAI, 2026; Xu et al., 2025b; Cui et al., 2026). These models increasingly act as conversational participants, yet no benchmark systematically evaluates whether they respond to what they hear and see with appropriate timing and content.

Engaging in such conversations requires two complementary capabilities. The first is spoken interaction, which requires interpreting turn-taking cues such as pauses and interruptions to decide when to speak and when to yield the floor. The second is visual understanding, which requires inferring the user's intent from facial expressions and actions and adjusting the response accordingly. The two must work together during the conversation.

However, current benchmarks address only part of this setting. Full-duplex speech benchmarks evaluate turn-taking behaviors such as interruption handling and backchanneling, but, being audio-only, they cannot account for how facial expressions and actions shape spoken behavior (Lin et al., 2025; Peng et al., 2025; Ge et al., 2025; Ray et al., 2026). Streaming video benchmarks treat the model as an observer and evaluate video comprehension and question answering (Lin et al., 2026c; Wang et al., 2025; Shi et al., 2026). In a video call, by contrast, the model takes part in the exchange as one of its two parties. The user's expressions and actions act as social cues that should regulate the model's turn timing, response content, and prosody. Voice-assistant benchmarks such as MULTIVOX evaluate whether responses to image or video inputs incorporate paralinguistic cues in

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.

the user's speech (Selvakumar et al., 2025). However, they adopt a single-turn question-answering format in which the visual input provides background context rather than depicting the user the model is conversing with. Because no existing benchmark measures spoken interaction and visual understanding together, when a model fails in audiovisual conversation, these benchmarks cannot determine whether the cause lies in (i) insufficient turn-taking skills, (ii) misinterpretation of the visual cue, or (iii) a failure to apply correct visual understanding during the conversation.

We introduce **OmniDuet-Bench**, a benchmark designed to disentangle these three sources of failure. The benchmark organizes full-duplex audiovisual conversation into three evaluation dimensions, each with two streaming tasks (Figure 1). **Interaction Capacity** examines turn-taking behavior: the model should yield the floor when interrupted but hold it through a backchannel, and wait through a within-turn pause but take the turn after a turn yield. **Visual Perception** probes the model's understanding of visual cues through direct questions: after the video ends, the model labels the user's facial expression, or describes the user's action and identifies its conversational function. **Cross-Modal Integration** examines whether the model uses these visual cues during conversation. The model receives no instruction to attend to expressions or actions and engages in natural conversation; we then assess whether the timing, content, and prosody of its response are consistent with the user's expression or action.

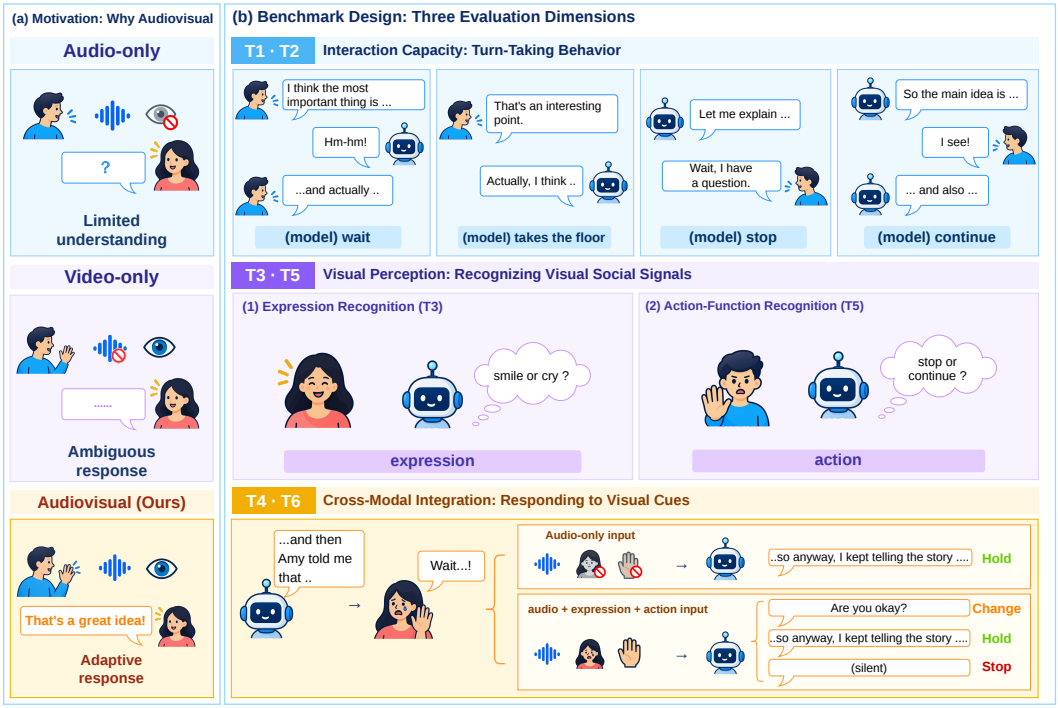


Figure 1: **OmniDuet-Bench tests whether a model acts on what it sees during conversation.** (a) Only audiovisual input supports an adaptive response. (b) The three evaluation dimensions: Interaction Capacity (T1, T2), Visual Perception (T3, T5), and Cross-Modal Integration (T4, T6).

Together, the three dimensions localize the source of a failure. A low Interaction Capacity score points to missing turn-taking skills, and a low Visual Perception score points to a failure to extract the required evidence from the video. When a model handles turn-taking well and recognizes the cue correctly but still responds inappropriately, the failure lies in Cross-Modal Integration; we call this failure to act on a correctly recognized cue the **perception-to-interaction gap**. To expose this gap, the Visual Perception and Cross-Modal Integration tasks share the same expression and action events, and the Cross-Modal Integration tasks additionally run under matched audiovisual and audio-only conditions. The benchmark comprises 707 human-annotated instances from natural two-person video calls (Agrawal et al., 2025), in which the model takes the place of one participant and receives only the user's audio and video in causal order.

We evaluate seven proprietary and open-source real-time omni-modal models and find that none leads on all three dimensions and that models often fail to hold the floor and to take the turn promptly when the user yields it, scoring far below the original speaker on interruption handling. Furthermore, Visual Perception remains limited, with low accuracy on both expressions and actions. Moreover, even when a model identifies an expression or action correctly, it often fails to adapt its response to the same cue during conversation, and removing the video does not lower most models' response scores, revealing a clear perception-to-interaction gap.

Our contributions are as follows:

1. **Evaluation framework (Section 3).** We decompose full-duplex audiovisual conversation into Interaction Capacity, Visual Perception, and Cross-Modal Integration. Six streaming tasks measure turn-taking behavior, visual recognition, and visually conditioned responses.
2. **Benchmark data (Section 3.1).** We curate 707 human-annotated task instances from natural two-person video calls. Recognition–response pairs on shared events and matched audiovisual and audio-only conditions separate perception failures from behavioral ones.
3. **Empirical findings (Section 4).** Across seven real-time omni-modal models, we find that no model leads on all three dimensions, that models often fail to hold the floor and to take the turn promptly when the user yields it, that Visual Perception remains limited, that recognizing a visual cue does not mean acting on it, and that video input does not consistently improve response quality.

2 Related Work

2.1 Omni-Modal Models

Omni-modal models process visual and audio inputs in a unified framework and generate text or speech. Qwen2.5-Omni and Qwen3-Omni couple multimodal perception with streaming speech generation (Xu et al., 2025a;b), VITA-1.5 supports near-real-time vision and speech interaction (Fu et al., 2025a), Stream-Omni enables image-grounded speech interaction (Zhang et al., 2025), and Ming-Omni unifies multimodal perception with speech and image generation (AI et al., 2025). More recent models extend these capabilities to multi-turn, proactive, and full-duplex interaction: InteractiveOmni targets audiovisual multi-turn dialogue (Tong et al., 2025), ROMA unifies proactive and reactive responses in streaming audiovisual understanding (Xueyun et al., 2026), and MiniCPM-o 4.5 and Qwen3.5-Omni support full-duplex real-time interaction (Cui et al., 2026; Team, 2026). These models provide the input and output capabilities that audiovisual conversation requires, but support for a modality does not imply that they adapt turn-taking and responses to visual cues.

2.2 Full-Duplex Spoken Dialogue Models

Full-duplex spoken dialogue models listen and speak simultaneously, handling overlapping speech from both parties. End-to-end models such as Moshi, SALMONN-omni, and PersonaPlex process input speech and generate output speech in parallel (Défossez et al., 2024; Yu et al., 2025; Roy et al., 2026), whereas modular systems add full-duplex control to half-duplex pipelines: FlexDuo uses a plug-and-play control module (Liao et al., 2025), FireRedChat pairs a turn controller with cascaded or semi-cascaded implementations (Chen et al., 2025), and DuplexCascade coordinates micro-turns with control tokens (Yang et al., 2026b). A few works incorporate visual input: Mini-Omni2 answers questions about visual content but relies on spoken commands for interruption (Xie & Wu, 2024), and AV-Dialog uses lip movements to track the target speaker and predict turn boundaries, while noting that it does not model facial expressions or gestures (Chen et al., 2026). None of these works evaluates whether facial expressions and gestures elicit appropriate conversational behavior.

2.3 Full-Duplex Dialogue Benchmarks

Full-duplex dialogue benchmarks evaluate both the timing of spoken interaction and the quality of responses. FD-Bench measures interruption handling, latency, and robustness with synthesized conversations and noise (Peng et al., 2025); Full-Duplex-Bench, its v1.5 version, and CoDeTT evaluate turn-taking, overlap handling, and context-dependent turn decisions (Lin et al., 2025; 2026b;

Shen et al., 2026); and Full-Duplex-Bench-v2 and MTR-DuplexBench extend evaluation to multi-turn conversation (Lin et al., 2026a; He et al., 2026). Other benchmarks examine state updates under interruption (EchoChain (Modi et al., 2026)), diverse interaction scenarios including model-initiated interruption (FLEXI (Ge et al., 2025)), emotional dialogue (HumDial (Zhao et al., 2026)), and task completion in service domains (τ-voice (Ray et al., 2026)). All of these benchmarks use audio-only input and therefore cannot assess how the user's expressions and actions affect the model's spoken behavior; OmniDuet-Bench retains their turn-level timing evaluation and adds visual cues from natural conversation.

2.4 Video and Audiovisual Understanding Benchmarks

Video and audiovisual understanding benchmarks evaluate how models understand and respond to streaming or offline media. StreamingBench and RTV-Bench examine continuous video understanding (Lin et al., 2026c; Xun et al., 2025), LongVALE and JointAVBench examine temporally grounded events and joint audiovisual reasoning (Geng et al., 2025; Chao et al., 2026), OmniMMI and LiViBench cover multimodal interaction and livestream understanding (Wang et al., 2025; 2026), and RIVER, PhoStream, and the OmniStar dataset introduced with LiveStar evaluate online or proactive responses (Shi et al., 2026; Lu et al., 2026; Yang et al., 2025). These benchmarks typically treat the model as an observer of the video rather than a participant in a conversation. Several works come closer to the conversational setting: in ViSpeak, people in the video address the model with scripted gesture instructions (Fu et al., 2025b); MULTIVOX evaluates whether responses reflect paralinguistic cues in the user's speech (Selvakumar et al., 2025); and SocialOmni uses multi-party clips to evaluate speaker identification and when a model acting as a designated participant should enter the conversation (Xie et al., 2026). None of them adopts two-person video calls or evaluates how the user's facial expressions shape the model's behavior. OmniDuet-Bench places the model in such a call and pairs Visual Perception and Cross-Modal Integration tasks on shared events to separate failures across the three dimensions.

3 OmniDuet-Bench

OmniDuet-Bench organizes its six tasks along the three evaluation dimensions introduced in Section 1. Table 1 lists the dimension of each task and its target output or behavior. By output format, the six tasks fall into two categories. **Recognition tasks** (T3, T5) pose a fixed question after playback ends, and the model answers with an expression label, or with an action description and its conversational function. **Spoken-behavior tasks** (T1, T2, T4, T6) track when the model starts speaking, waits, continues, stops, or adapts its response while the media streams. Recognition tasks constitute the Visual Perception dimension; spoken-behavior tasks constitute the other two. We next describe the data (Section 3.1), tasks (Section 3.2), evaluation pipeline (Section 3.3), and scoring (Section 3.4).

Table 1: Evaluation dimensions, tasks, target outputs, and instance counts of OmniDuet-Bench.

	Task	Target output or behavior	Instances
Interaction Capacity			
T1	Interruption handling	Stop for interruptions; continue through backchannels	179
T2	Turn entry	Wait through pauses; respond after turn yields	194
Visual Perception			
T3	Expression recognition	One of six facial-expression labels	99
T5	Action-function recognition	Action description and its conversational function	68
Cross-Modal Integration			
T4	Expression-conditioned response	Response appropriate to the user's expression	99
T6	Action-conditioned response	Stop, adapt, or continue according to the action	68

3.1 Dataset Construction and Annotation

OmniDuet-Bench curates samples from natural two-person video calls in the Seamless Interaction corpus (Agrawal et al., 2025). Participants hold unscripted remote video conversations recorded locally on each side, so their expressions and actions are spontaneous. Compared with synthetic dialogues or scripted videos, natural dyadic conversations yield interaction events closer to those full-duplex models encounter in deployment. Construction has three stages (Figure 2): automatic candidate extraction from the corpus's precomputed metadata, human review and correction, and task construction. The final benchmark draws on 440 distinct conversations involving 233 speakers.

Candidate extraction and human annotation. In the first stage, we extract candidate events automatically. Turn events come from voice activity segments and word-level transcripts, expression events from per-frame emotion-classification logits, and action events from 30 Hz body kinematic trajectories. Three annotators independently review each candidate through a web-based annotation interface, and a majority vote decides the final label. A fourth senior annotator arbitrates when no majority exists. Appendix F describes the extraction rules, annotation interface, and quality control.

The benchmark contains 707 task instances. Table 1 gives per-task instance counts, and Table 10 (Appendix F) gives class distributions.

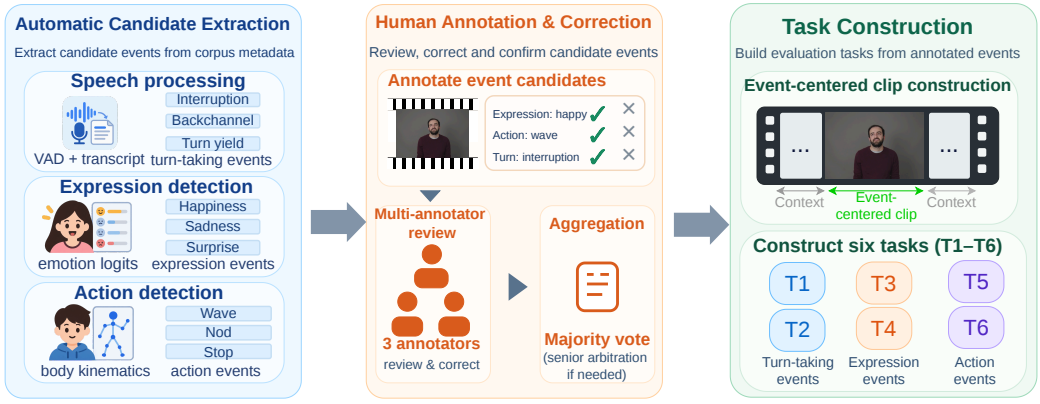


Figure 2: **Data construction pipeline.** Candidate extraction, human annotation, and task construction.

Clip construction. For each sample, we extract a media clip centered on the annotated event that retains the preceding dialogue and the subsequent interaction. Appendix E.1 gives the clipping rules and clip durations for each task. Clips for a recognition task (T3 or T5) and its paired response task (T4 or T6) share the event participant, the temporal boundaries, and the dialogue context before the event. The recognition clip ends with the target event, whereas the response clip continues past it.

3.2 Task Design

T1: Interruption Handling (Interaction Capacity). The model receives a role prompt generated from the pre-event dialogue history and continues speaking in place of one participant. While the model speaks, the audiovisual input contains an event annotated as either an interruption or a backchannel. The model should yield the floor when interrupted and hold it through backchannels. This task evaluates whether the model distinguishes these two events.

T2: Turn Entry (Interaction Capacity). The model receives a fixed role prompt and participates as the listener. The audiovisual input contains an event annotated as either a within-turn pause or a turn yield. The model should remain silent during pauses and respond promptly after turn yields. This task evaluates whether the model distinguishes a within-turn pause from a turn yield.

T3: Expression Recognition (Visual Perception). The model receives a video clip containing an annotated expression event and, after playback, selects one label from six classes: `ANGER`, `CONTEMPT`, `DISGUST`, `HAPPINESS`, `SADNESS`, `SURPRISE_OR_FEAR`.

T4: Expression-Conditioned Response (Cross-Modal Integration). The model receives a video clip of the same event as T3, but the clip continues after the expression appears. The model responds as in natural conversation, and we assess whether its response is consistent with the user's expression and the dialogue context.

T5: Action-Function Recognition (Visual Perception). The model receives a video clip containing an annotated action event. After playback, it describes the observed body movement and classifies its conversational function into one of three classes: `INTERRUPTS_DIALOGUE` (asks the other participant to stop or yield the floor), `ALTERS_TONE_OR_MEANING` (signals that the other participant should adapt how they respond), and `NO_CONVERSATIONAL_EFFECT` (does not require the other participant to change behavior). This task evaluates whether the model understands both the action content and its conversational function.

T6: Action-Conditioned Response (Cross-Modal Integration). The model receives a video clip of the same event as T5 and actively speaks in place of one participant. During its speech, the video input contains an annotated action event. The model should stop, adapt, or continue its current speech according to the action's conversational function.

Shared events and matched conditions. T3 and T4 share 99 expression events; T5 and T6 share 68 action events. For each shared event, the Visual Perception task (T3 or T5) and the Cross-Modal Integration task (T4 or T6) run independently and receive separate scores. Pairing the recognition result with the response result for the same event distinguishes a failure to recognize the cue from a failure to act on a correctly recognized cue.

In addition, T4 and T6 each run twice, once with audio and video and once with audio only, with identical audio, prompts, and event timing. The per-sample score difference therefore measures what the video contributes to the response. Combined with the recognition–response pairing, this identifies cases in which a model perceives a cue but does not act on it: the model identifies the expression or action in the Visual Perception task, yet adding the video does not raise its response score.

3.3 Evaluation Pipeline

As introduced in Section 1, the model takes the place of one participant and receives the user's audio and video in causal order. The benchmark hides the recorded responses of the replaced participant (the original speaker) from the model and uses them only as scoring references. The benchmark builds a task prompt for each sample and records the model output. Figure 3 summarizes the full process from input construction to scoring and paired analysis.

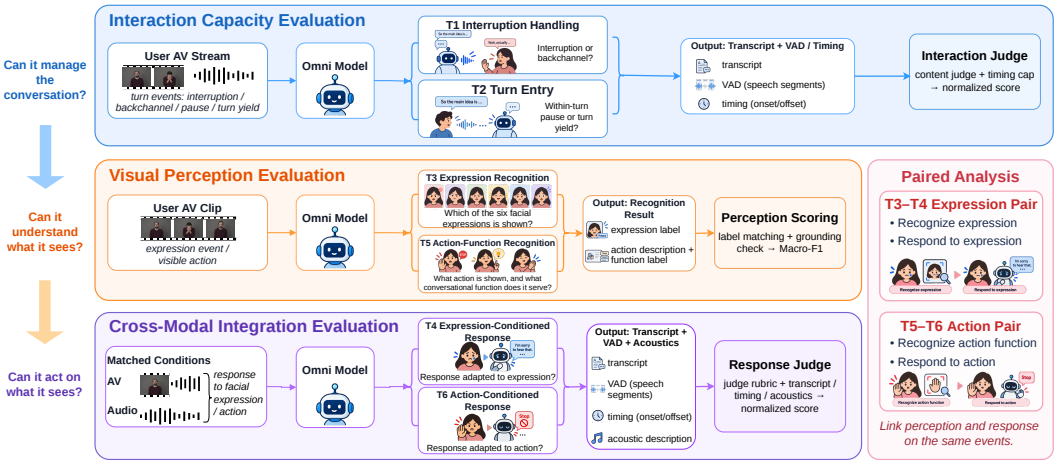


Figure 3: **Evaluation pipeline.** Each dimension streams the audiovisual input to the model and scores its output with dimension-specific evidence; the benchmark then pairs recognition and response results on shared events.

After each run, the scorer combines evidence from the model output with the reference annotations to compute the task score (Section 3.4); the benchmark then pairs recognition and response results on shared events, and audiovisual and audio-only results on the same samples (Section 3.2).

3.4 Evaluation Protocol and Scoring

Recognition tasks compare the model's answers with the human annotations. Spoken-behavior tasks assess both turn timing and response content: we compute timing from the model's speech intervals, and a judge model evaluates the content and prosody of the response.

Recognition task scoring. T3 computes macro-F1 over the six expression labels. T5 requires the model to output both an action description and a function label, so its scoring uses two layers to ensure that function judgments rest on correct action understanding. In the first layer, the judge model Qwen3.8-Max (Qwen Team, 2026) verifies the semantic consistency between the action description and the human reference on a 0–2 scale (0 = unrelated or contradictory, 1 = only a body region or generic motion identified, 2 = core action correctly identified). The second layer computes macro-F1 over the three function labels, but treats the function prediction as incorrect for every sample whose action description fails verification (score < 2), even when the label happens to be correct. This grounded macro-F1 serves as the primary T5 metric.

Spoken-behavior task scoring. We reconstruct all evidence for spoken-behavior tasks from the audio actually played back, not from the text the model returns. Qwen3-ASR-1.7B transcribes the output audio, and Qwen3-ForcedAligner-0.6B provides word-level timestamps. A voice activity detector (VAD, Appendix E.2) locates speech segments $\{(s_i, e_i)\}$, from which we compute stop latency, reaction latency, and speaking status. T4 and T6 also use MOSS-Audio-4B-Instruct (Yang et al., 2026a) to describe vocal delivery, covering intonation, loudness, speaking rate, pauses, and nonverbal vocalizations such as laughter and sighs.

T4 evaluates only the model's first response after the expression appears; T6 evaluates a bounded window from 3 s before action onset to the end of observation. Appendix B.3 gives the T4 first-response selection rule.

The judge model is Qwen3.8-Max (Qwen Team, 2026) (fixed snapshot, non-thinking mode, temperature = 0), and timing and content are judged separately. For T1, T2, and T6 stop samples, a program determines the timing outcome from speech intervals, and the judge scores only response content from 0 to 4 without seeing the event class, timing outcome, or latency. T4 and the T6 adapt and continue samples have no fixed deadline. For these samples, the judge combines the event class, dialogue context, transcript, speech intervals, and acoustic description to output $r \in \{0, 1, 2, 3, 4\}$. A score of 0 means the target behavior is absent or clearly inappropriate, and 4 means the response is timely, specific, and preserves the conversational thread. Appendix B gives the rubrics; the judge agrees with two other judges ($\kappa = 0.78$–0.84), and its scores pass manual review (Appendix B.8).

Timing caps. For T1, T2, and T6 stop samples, the program sets a score cap from the timing outcome, and the final score is the smaller of the content score and the cap. The cap is 0 when the model is silent at an event where it should be speaking, never stops when it should yield, never starts when it should take the turn, starts speaking before the user yields the turn, or stops after a backchannel without resuming. The cap is 1 when the model exhibits the correct behavior but misses the deadline, speaks during a pause, or stops after a backchannel and then resumes. Correct timing imposes no cap, and a T2 pause sample in which the model stays silent receives the full score of 4 without a judge call. T6 continue samples score 0 when the model never responds after the action. Appendix B.7 gives all rules and thresholds.

Task score aggregation. We normalize each sample's final score r' to $[0, 1]$ as $\hat{r} = r'/4$, where $r' = r$ for samples without a timing cap. Each task groups samples by class label $l \in L$, computes the within-group mean, and then takes the macro average over groups:

$$S_{\text{task}} = \frac{1}{|L|} \sum_{l \in L} \frac{1}{|R_l|} \sum_{i \in R_l} \hat{r}_i \tag{1}$$

Recognition tasks use macro-F1 directly as the primary metric. All primary metrics range over $[0, 1]$, and we report them multiplied by 100. Failed runs count as zero and remain in the denominator.

4 Experiments and Results

We evaluate seven models that support real-time audiovisual streaming interaction. Four are proprietary API models: Gemini 3.8 Live (Google Cloud, 2026), Gemini 3.1 Flash Live Preview (Google Cloud, 2026), GPT-Realtime-2.1 (OpenAI, 2026), and Qwen3.5-Omni-Plus-Realtime (Team, 2026). One is a lightweight API model, Qwen-Omni-Turbo-Realtime (Xu et al., 2025a), and two are open-source models, MiniCPM-o 4.5 (9B) (Cui et al., 2026) and VITA-1.5 (Fu et al., 2025a). All models share one input protocol: 1 fps video frames and 16 kHz mono audio are streamed in one-second chunks in causal order, preventing access to future content. Each sample runs once per condition, and T4 and T6 run again under the audio-only condition. Table 2 reports the scores of all models on the six tasks. As a human reference, we score the original speaker's recordings on T1 with the same pipeline (Appendix G). Appendix D gives access methods, inference parameters, evaluation scale, and tool models, Appendix H gives per-class results, and Appendix J presents case studies.

Table 2: Main results of seven models on six tasks (0–100). Parentheses give the T4 and T6 scores under the audio-only condition; all other scores, including Avg., use audiovisual input. T1 timing reports the percentage of all interruption samples with a stop within 1.5 s and the median stop latency (s). Bold marks the best audiovisual model score in each column.

Model *Dimension*	**T1** *Inter.*	**T2** *Inter.*	**T3** *Perc.*	**T4** *Integ.*	**T5** *Perc.*	**T6** *Integ.*	**Avg.**	**T1 timing (% / s)**
Original speaker	82.2	–	–	–	–	–	–	80 / 0.40
Qwen3.5-Omni-Plus-Realtime	27.0	39.7	33.1	**71.4** (70.4)	**47.9**	**51.8** (49.7)	**45.1**	38 / 0.43
Gemini 3.1 Flash Live Preview	23.2	**61.0**	27.5	60.9 (65.6)	44.7	34.3 (33.9)	41.9	23 / 1.04
Gemini 3.8 Live	17.1	58.6	29.8	59.1 (64.1)	32.8	31.2 (38.4)	38.1	32 / 0.44
GPT-Realtime-2.1	10.4	57.7	**35.6**	41.8 (50.9)	32.5	14.8 (18.5)	32.1	15 / 0.41
MiniCPM-o 4.5	**29.6**	55.0	19.7	44.9 (15.3)	10.8	29.3 (17.3)	31.6	12 / 3.93
Qwen-Omni-Turbo-Realtime	8.6	49.4	33.6	40.5 (48.9)	3.6	27.4 (31.8)	27.2	23 / 1.02
VITA-1.5	18.0	23.7	24.5	47.7 (53.0)	5.7	32.4 (35.4)	25.3	14 / 3.36

Finding 1: No model leads on all three dimensions. The best six-task average, from Qwen3.5-Omni-Plus-Realtime, is only 45.1, and each model's strengths shift across dimensions. MiniCPM-o 4.5 has the highest Interaction Capacity average, whereas Qwen3.5-Omni-Plus-Realtime leads on Visual Perception and Cross-Modal Integration. GPT-Realtime-2.1 scores highest on T3 (35.6) but lowest on T6 (14.8); MiniCPM-o 4.5 scores highest on T1 (29.6) but lowest on T3 (19.7); and Qwen3.5-Omni-Plus-Realtime leads on T4, T5, and T6 yet ranks sixth on T2 (39.7). Strength in one dimension thus does not carry over to the others.

Finding 2: Models often fail to hold the floor and to take the turn promptly when the user yields it. Every model scores at most 29.6 on T1, compared with 82.2 for the original speaker. The gap stems mainly from a failure to hold the floor: when the interruption arrives, models are speaking in only 19%–55% of samples. When they are speaking, most models stop in time: Qwen3.5-Omni-Plus-Realtime, Gemini 3.8 Live, GPT-Realtime-2.1, and Qwen-Omni-Turbo-Realtime stop within 1.5 s in 80%–100% of such cases, on par with the original speaker's 80%, and the first three have median stop latencies of 0.41–0.44 s, close to the original speaker's 0.40 s; MiniCPM-o 4.5 and VITA-1.5 do so in only 21% and 25%. On T2, no model takes the turn within the deadline after more than 30% of turn yields. Behavior at within-turn pauses is polarized: GPT-Realtime-2.1 stays silent through 99% of pauses yet never takes the turn within the deadline after a turn yield, while VITA-1.5 and Qwen3.5-Omni-Plus-Realtime stay silent through only 15% and 27% of pauses. The timing problems of current models thus lie in holding the floor and taking the turn at the right moment.

Table 3: Recognition–response pairing on the 99 expression events (T3/T4) and 68 action events (T5/T6). The first two columns of each pair give the percentage of appropriate responses when recognition is correct and when it is wrong; the third (Gap) gives the number of events recognized correctly but answered inappropriately.

	T3/T4 (expressions)			T5/T6 (actions)		
Model	**Correct**	**Wrong**	**Gap**	**Correct**	**Wrong**	**Gap**
Qwen3.5-Omni-Plus-Realtime	80.0	61.1	9	64.3	40.0	10
Gemini 3.1 Flash Live Preview	61.5	50.0	15	34.8	31.1	15
Gemini 3.8 Live	59.0	41.7	16	12.5	28.8	14
GPT-Realtime-2.1	39.1	32.1	28	6.7	13.2	14
MiniCPM-o 4.5	41.7	41.3	21	33.3	29.0	4
Qwen-Omni-Turbo-Realtime	32.7	21.3	35	50.0	19.7	1
VITA-1.5	31.4	25.0	24	0.0	21.5	3

Finding 3: Recognizing a visual cue does not mean acting on it. Visual Perception itself is limited: the best scores are 35.6 on T3 and 47.9 on T5. Table 3 pairs the recognition and response results on the same events, counting a response as appropriate when it scores at least 3 of 4 and a T5 recognition as correct only when both the action description and the function label are correct. On expression events, correct recognition raises the rate of appropriate responses by only 0.4–18.9 percentage points, and each model still responds inappropriately to 9–35 of the expressions it recognizes correctly. On action events, the link between recognition and response disappears: Gemini 3.8 Live and GPT-Realtime-2.1 respond appropriately less often after identifying the action and its function correctly (12.5% and 6.7%) than after misidentifying it (28.8% and 13.2%). No model scores above 33.3 on T6 stop samples: models rarely stop promptly after seeing an interrupting action. This is the perception-to-interaction gap: models can recognize a visual cue yet do not carry it into the conversation.

Finding 4: Video input does not consistently improve response quality. The parenthesized scores in Table 2 report the same samples under the audio-only condition. Without video, five of the seven models score higher on T4 and four score higher on T6; GPT-Realtime-2.1, for example, gains 9.1 points on T4. Only Qwen3.5-Omni-Plus-Realtime (1.0 points higher on T4 and 2.1 on T6) and MiniCPM-o 4.5 score higher on both tasks with video. At the sample level, among the events a model recognizes correctly in the Visual Perception task, adding the video fails to raise the response score for 76% of expressions and 85% of actions. Together with Finding 3, these results show that current models make little use of the user's expressions and actions to adapt their responses.

5 Conclusion

OmniDuet-Bench evaluates full-duplex audiovisual conversation along Interaction Capacity, Visual Perception, and Cross-Modal Integration, using six streaming tasks built from 707 task instances drawn from natural two-person video calls. Our evaluation of seven real-time omni-modal models shows that no model leads on all three dimensions and finds failures of all three kinds: models often fail to hold the floor and to take the turn promptly (Interaction Capacity), recognize expressions and actions with limited accuracy (Visual Perception), and often do not act on cues they recognize correctly (Cross-Modal Integration). Recognition–response pairing shows that recognizing an expression or action does not mean acting on it, and removing the video does not lower most models' response scores. These findings indicate that existing models do not yet translate visual understanding into conversational behavior, and future models need to close this perception-to-interaction gap. We will release the benchmark data and evaluation code under the license of the source corpus. Appendix I discusses limitations and future work.

AI use statement

In this work, we used generative AI tools to aid and polish the writing of this paper, for retrieval and discovery when compiling related work, and for research ideation and execution, including

assistance in refining the evaluation protocol and in writing and testing the evaluation-harness code. We have not used generative AI tools to draft sections of the paper, to generate synthetic datasets, or to prove mathematical claims. We have reviewed all AI-assisted work: the authors verified every technical claim in the manuscript against the protocol implementation and run records, checked AI-retrieved related work against the original papers, and verified AI-assisted code through unit tests and manual audits of evaluation outputs. Separately, large language models also operate inside the benchmark as the response judge, the acoustic-description extractor, and the dialogue-history summarizer. They are components of the evaluation protocol, fully specified in Appendices B, C, and D, not writing tools. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

This work collects no new human data. All samples come from the publicly released Seamless Interaction corpus (Agrawal et al., 2025), which is distributed under the CC-BY-NC 4.0 license; we use it only for non-commercial research and will release the derived benchmark data under the same license. The corpus contains participants' faces and voices; we add no identity information and create no annotations related to identity recognition. Expression labels describe observable facial expressions and do not represent the participants' internal emotional states. Annotators complete practice samples before beginning annotation (Appendix F.3). Conversational models that read users' expressions and actions could be misused for surveillance or emotional manipulation; this benchmark evaluates only whether models respond appropriately to these cues in conversation and provides no tool for inferring an individual's emotions or intentions.

Reproducibility statement

We will release the benchmark manifest, the media clips under the source corpus license, each model's adapter and evaluation code, and the judge model's per-item outputs. The paper and appendix give every setting needed to reproduce the evaluation. Appendix A gives all prompts sent to the models; Appendix B gives the judge rubrics, judge input fields, and the formal definition of the timing rules; Appendix C gives the acoustic-description prompts; Appendix D gives model access, the input protocol, inference parameters, and each model's adapter parameters; Appendix E gives the clip extraction rules and voice activity detection parameters; and Appendix F gives the candidate extraction and annotation procedure. The transcription model Qwen3-ASR-1.7B and the acoustic-description model MOSS-Audio-4B-Instruct are open-weight models; we call the judge model Qwen3.8-Max through an API with the fixed snapshot qwen3.8-max-0902 and release its per-item outputs. Both the judge and the acoustic description use greedy decoding (temperature = 0). We call proprietary models through vendor APIs, and their results reflect the versions we accessed in September 2026.

References

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, Praveen Chowdary, Joe Chuang, Antony D'Avirro, Jon Daly, Ning Dong, Mark Duppenthaler, Cynthia Gao, Jeff Girard, Martin Gleize, Sahir Gomez, Hongyu Gong, Srivathsan Govindarajan, Brandon Han, Sen He, Denise Hernandez, Yordan Hristov, Rongjie Huang, Hirofumi Inaguma, Somya Jain, Raj Janardhan, Qingyao Jia, Christopher Klaiber, Dejan Kovachev, Moneish Kumar, Hang Li, Yilei Li, Pavel Litvin, Wei Liu, Guangyao Ma, Jing Ma, Martin Ma, Xutai Ma, Lucas Mantovani, Sagar Miglani, Sreyas Mohan, Louis-Philippe Morency, Evonne Ng, Kam-Woh Ng, Tu Anh Nguyen, Amia Oberai, Benjamin Peloquin, Juan Pino, Jovan Popovic, Omid Poursaeed, Fabian Prada, Alice Rakotoarison, Rakesh Ranjan, Alexander Richard, Christophe Ropers, Safiyyah Saleem, Vasu Sharma, Alex Shcherbyna, Jia Shen, Jie Shen, Anastasis Stathopoulos, Anna Sun, Paden Tomasello, Tuan Tran, Arina Turkatenko, Bo Wan, Chao Wang, Jeff Wang, Mary Williamson, Carleigh Wood, Tao Xiang, Yilin Yang, Julien Yao, Chen Zhang, Jiemin Zhang, Xinyue Zhang, Jason Zheng, Pavlo Zhyzheria, Jan Zikes, and Michael Zollhoefer. Seamless Interaction: Dyadic audiovisual motion modeling and large-scale dataset, 2025. URL https://arxiv.org/abs/2506.22554.

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, Furong Xu, GuangMing Yao, Jun Zhou, Jingdong Chen, Jianxin Sun, Jiajia Liu, Jianjiang Zhu, Jun Peng, Kaixiang Ji, Kaiyou Song, Kaimeng Ren, Libin Wang, Lixiang Ru, Lele Xie, Longhua Tan, Lyuxin Xue, Lan Wang, Mochen Bai, Ning Gao, Pei Chen, Qingpei Guo, Qinglong Zhang, Qiang Xu, Rui Liu, Ruijie Xiong, Sirui Gao, Tinghao Liu, Taisong Li, Weilong Chai, Xinyu Xiao, Xiaomei Wang, Xiaoxue Chen, Xiao Lu, Xiaoyu Li, Xingning Dong, Xuzheng Yu, Yi Yuan, Yuting Gao, Yunxiao Sun, Yipeng Chen, Yifei Wu, Yongjie Lyu, Ziping Ma, Zipeng Feng, Zhijiang Fang, Zhihao Qiu, Ziyuan Huang, and Zhengyu He. Ming-Omni: A unified multimodal model for perception and generation, 2025. URL https://arxiv.org/abs/2506.09344.

Jianghan Chao, jianzhang gao, Wenhui Tan, Yuchong Sun, Ruihua Song, and Liyun Ru. JointAVBench: A benchmark for joint audio-visual reasoning evaluation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Zg1YH8R5GG.

Junjie Chen, Yao Hu, Junjie Li, Kangyue Li, Kun Liu, Wenpeng Li, Xu Li, Ziyuan Li, Feiyu Shen, Xu Tang, Manzhen Wei, Yichen Wu, Fenglong Xie, Kaituo Xu, and Kun Xie. FireRedChat: A pluggable, full-duplex voice interaction system with cascaded and semi-cascaded implementations, 2025. URL https://arxiv.org/abs/2509.06502.

Tuochao Chen, Bandhav Veluri, Hongyu Gong, and Shyamnath Gollakota. AV-Dialog: Spoken dialogue models with audio-visual input. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42208–42225, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1954. URL https://aclanthology.org/2026.acl-long.1954/.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction, 2026. URL https://arxiv.org/abs/2604.27393.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue, 2024. URL https://arxiv.org/abs/2410.00037.

Chaoyou Fu, Haojia Lin, Xiong Wang, yifan zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long MA, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. VITA-1.5: Towards GPT-4o level real-time vision and speech interaction. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 75300–75320. Curran Associates, Inc., 2025a. doi: 10.52202/085713-2529. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6ce9d51dded7dac82b3b4d3dbb1d73bc-Paper-Conference.pdf.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Yi-Xing Peng, Kun-Yu Lin, Xihan Wei, Jian-Fang Hu, Xiaohua Xie, and Wei-Shi Zheng. ViSpeak: Visual instruction feedback in streaming videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21778–21788, 2025b. doi: 10.1109/ICCV51701.2025.02022.

Yuan Ge, Saihan Chen, Jingqi Xiao, Xiaoqian Liu, Tong Xiao, Yan Xiang, Zhengtao Yu, and Jingbo Zhu. FLEXI: Benchmarking full-duplex Human-LLM speech interaction, 2025. URL https://arxiv.org/abs/2509.22243.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. LongVALE: Vision-audio-language-event benchmark towards time-aware omni-modal perception of

long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18959–18969, 2025. doi: 10.1109/CVPR52734.2025.01766.

Google Cloud. Gemini live API overview. Google Cloud Documentation, 2026. URL https://cloud.google.com/vertex-ai/generative-ai/docs/live-api. Accessed 2026-09-16.

Zhang He, Wenqian Cui, Haoning Xu, Xiao-Hui Li, Lei Zhu, Haoli Bai, Ma Shaohua, and Irwin King. MTR-DuplexBench: Towards a comprehensive evaluation of multi-round conversations for full-duplex speech language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5334–5351, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.263. URL https://aclanthology.org/2026.findings-acl.263/.

Borui Liao, Yulong Xu, Jiao Ou, Kaiyuan Yang, Weihua Jian, Pengfei Wan, and Di Zhang. FlexDuo: A pluggable system for enabling full-duplex capabilities in speech dialogue systems, 2025. URL https://arxiv.org/abs/2502.13472.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu, and Hung-Yi Lee. Full-Duplex-Bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2025. doi: 10.1109/ASRU65441.2025.11433838.

Guan-Ting Lin, Shih-Yun Shan Kuan, Jiatong Shi, Kai-Wei Chang, Siddhant Arora, Shinji Watanabe, and Hung-yi Lee. Full-Duplex-Bench-v2: A multi-turn evaluation framework for duplex dialogue systems with an automated examiner. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 27–36, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-391-3. doi: 10.18653/v1/2026.acl-short.4. URL https://aclanthology.org/2026.acl-short.4/.

Guan-Ting Lin, Shih-Yun Shan Kuan, Qirui Wang, Jiachen Lian, Tingle Li, Shinji Watanabe, and Hung-Yi Lee. Full-Duplex-Bench v1.5: Evaluating overlap handling for full-duplex speech models. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 19447–19451, 2026b. doi: 10.1109/ICASSP55912.2026.11463576.

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, 2026c. doi: 10.1109/ICASSP55912.2026.11463959.

Xudong Lu, Guan Huankang, Yang Bo, Jinpeng Chen, Xintong Guo, Shuhan LI, Fang Liu, Peiwen Sun, Xueying Li, Wei Zhang, Xue Yang, Rui Liu, and Hongsheng Li. PhoStream: Benchmarking real-world streaming for omnimodal assistants in mobile scenarios. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=cCMbxJcDmH.

Smit Nautambhai Modi, Gandharv Mahajan, Marc Wetter, and Randall Welles. EchoChain: A full-duplex benchmark for state-update reasoning under interruptions, 2026. URL https://arxiv.org/abs/2604.16456.

OpenAI. Realtime API. OpenAI API Documentation, 2026. URL https://platform.openai.com/docs/guides/realtime. Accessed 2026-09-16.

Yizhou Peng, Yi-Wen Chao, Dianwen Ng, Yukun Ma, Chongjia Ni, Bin Ma, and Eng Siong Chng. FD-Bench: A Full-Duplex Benchmarking Pipeline Designed for Full Duplex Spoken Dialogue Systems. In *Interspeech 2025*, pp. 176–180, 2025. doi: 10.21437/Interspeech.2025-739.

Qwen Team. Qwen3.8-Max (snapshot qwen3.8-max-0902). Alibaba Cloud Model Studio, 2026.

Soham Ray, Keshav Dhandhania, Victor Barres, and Karthik R Narasimhan. τ-Voice: Benchmarking full-duplex voice agents on real-world domains. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=2Oj6fg0m1j.

Rajarshi Roy, Jonathan Raiman, Sang-Gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. PersonaPlex: Voice and role control for full duplex conversational speech models. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16137–16141, 2026. doi: 10.1109/ICASSP55912.2026.11463413.

Ramaneswaran Selvakumar, Ashish Seth, Nishit Anand, Utkarsh Tyagi, Sonal Kumar, Sreyan Ghosh, and Dinesh Manocha. MULTIVOX: A benchmark for evaluating voice assistants for multimodal interactions. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28481–28493, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1447. URL https://aclanthology.org/2025.emnlp-main.1447/.

Huan Shen, Yingao Wang, Shangkun Huang, Wei Zou, and Yunzhang Chen. CoDeTT: A context-aware decision benchmark for turn-taking evaluation, 2026. URL https://arxiv.org/abs/2603.25434.

Yansong Shi, Qingsong Zhao, Tianxiang Jiang, Xiangyu Zeng, Yi Wang, and Limin Wang. RIVER: A real-time interaction benchmark for video LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xmtvHH62Ic.

Qwen Team. Qwen3.5-Omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, Xueheng Li, Lumin Li, Chenxu Guo, Jiasheng Zhou, Jiandong Chen, Xianye Wu, Jiahao Wang, Silei Wu, Lei Chen, Hanming Deng, Yuxuan Song, Dinghao Zhou, Guiping Zhong, Ken Zheng, Shiyin Kang, and Lewei Lu. InteractiveOmni: A unified omni-modal model for audio-visual multi-turn dialogue, 2025. URL https://arxiv.org/abs/2510.13747.

Xiaodong Wang, Langling Huang, Zhirong Wu, Xu Zhao, Teng Xu, Xuhong Xia, and Peixi Peng. LiViBench: An omnimodal benchmark for interactive livestream video understanding, 2026. URL https://arxiv.org/abs/2601.15016.

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, 2025. doi: 10.1109/CVPR52734.2025.01763.

Tianyu Xie, Jinfa Huang, Yuexiao Ma, Rongfang Luo, Yan Yang, Qingchuan Ma, Wang Chen, Yuhui Zeng, Yixuan Zou, Zhiqiang Lu, Ruize Fang, Jiebo Luo, Rongrong Ji, and Xiawu Zheng. SocialOmni: Benchmarking audio-visual social interactivity in omni models, 2026. URL https://arxiv.org/abs/2603.16859.

Zhifei Xie and Changqiao Wu. Mini-Omni2: Towards open-source GPT-4o with vision, speech and duplex capabilities, 2024. URL https://arxiv.org/abs/2410.11190.

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, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report, 2025b. URL https://arxiv.org/abs/2509.17765.

Tian Xueyun, Wei Li, Bingbing Xu, Heng Dong, Yuanzhuo Wang, and Huawei Shen. ROMA: Real-time omni-multimodal assistant with interactive streaming understanding. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23018–23039, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1153. URL https://aclanthology.org/2026.findings-acl.1153/.

ShuHang Xun, Sicheng Tao, Jungang Li, Yibo Shi, Zhixin Lin, Zhanhui Zhu, Yibo Yan, Hanqian Li, LingHao Zhang, Shikang Wang, Yixin Liu, Hanbo Zhang, Ying Ma, and Xuming Hu. RTV-Bench: Benchmarking MLLM continuous perception, understanding and reasoning through real-time video. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-0600. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/19e4ea30dded58259665db375885e412-Paper-Datasets_and_Benchmarks_Track.pdf.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, Zhengyuan Lin, Ziqi Chen, Zhaoye Fei, Chenghao Liu, Donghua Yu, Jun Zhan, Kang Yu, Kexin Huang, Liwei Fan, Mingshu Chen, Qinyuan Cheng, Ruixiao Li, Shimin Li, Songlin Wang, Xingjian Zhao, Yang Gao, Yitian Gong, Yiyang Zhang, Zhe Xu, and Xipeng Qiu. MOSS-Audio technical report, 2026a. URL https://arxiv.org/abs/2606.01802.

Jianing Yang, Yusuke Fujita, and Yui Sudo. DuplexCascade: Full-duplex speech-to-speech dialogue with VAD-Free cascaded ASR-LLM-TTS pipeline and micro-turn optimization, 2026b. URL https://arxiv.org/abs/2603.09180.

Zhenyu Yang, Kairui Zhang, Yuhang Hu, Bing Wang, Shengsheng Qian, Bin Wen, Fan Yang, Tingting Gao, Weiming Dong, and Changsheng Xu. LiveStar: Live streaming assistant for real-world online video understanding. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 31266–31304. Curran Associates, Inc., 2025. doi: 10.52202/085713-1050. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2ce4f0b8e24c45318352068603153590-Paper-Conference.pdf.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. SALMONN-omni: A standalone speech LLM without codec injection for full-duplex conversation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 24617–24643. Curran Associates, Inc., 2025. doi: 10.52202/085713-0829. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/233aee920dab065709145371b5900b8f-Paper-Conference.pdf.

Shaolei Zhang, Shoutao Guo, Qingkai Fang, Yan Zhou, and Yang Feng. Stream-Omni: Simultaneous multimodal interactions with large language-vision-speech model, 2025. URL https://arxiv.org/abs/2506.13642.

Zhixian Zhao, Shuiyuan Wang, Guojian Li, Hongfei Xue, Chengyou Wang, Shuai Wang, Longshuai Xiao, Zihan Zhang, Hui Bu, Xin Xu, Xinsheng Wang, Hexin Liu, Eng Siong Chng, Hung-Yi Lee, and Lei Xie. The ICASSP 2026 HumDial challenge: Benchmarking human-like spoken dialogue systems in the LLM era. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21850–21852, 2026. doi: 10.1109/ICASSP55912.2026.11460911.

A TASK PROMPTS

Table 4 lists the system prompt and post-prompt for each task. We show the prompts in English exactly as sent to the models. The T1 system prompt contains the replaced speaker's last utterance before the takeover point, and the T6 system prompt contains dynamic dialogue-history summaries; the other tasks use fixed prompts.

Table 4: Task prompts sent to the evaluated models. {`utterance`} is the replaced speaker's last utterance before the takeover point, taken verbatim from the source transcript; {`earlier`} and {`recent`} are dynamically filled dialogue-history summaries. For T3 and T5, the part below the dashed line is the post-prompt.

T1 system prompt

You are a helpful assistant participating naturally in a video call. Continue the current discussion without mentioning these instructions. The point you were explaining immediately before this clip was: "{`utterance`}". Start speaking now to continue the conversation.

T2 system prompt

You are the other participant in a natural realtime video conversation. Listen to and watch the person continuously. Respond naturally when they have finished and offered you the floor. During a pause within their unfinished turn, keep listening without speaking. Do not describe the video or explain these instructions.

T3 system prompt

You are completing a closed-set facial-expression recognition evaluation. Watch the entire video and remain silent until the final question. Treat the video solely as evidence for the final question.

Post-prompt: Focus only on the visible person's facial expression during the final 1.0 second of the video. Choose the single most prominent expression from: ANGER, CONTEMPT, DISGUST, HAPPINESS, SADNESS, SURPRISE_OR_FEAR. Return only the exact label and no other words.

T4 system prompt

You are the other participant in a natural realtime video conversation. Watch and listen continuously, and respond as the conversation unfolds. Let the person's visible expression, spoken content, and vocal tone guide your wording and delivery. Respond naturally. Describe the video or name an emotion only when doing so would be natural in the conversation.

T5 system prompt

You are completing a closed-set video understanding task. Watch and listen to the entire clip before answering. Focus on the visible person's final action and its conversational function in context. Do not participate in the conversation, describe your reasoning, or answer before playback ends. After playback, answer in exactly two lines: ACTION: one short literal description of the visible body movement or nonverbal behavior LABEL: one exact label from the final question. Do not add any other text.

Post-prompt: Answer in exactly two lines: ACTION: ... LABEL: ... Describe the visible person's final action, then classify its conversational function. For the LABEL field, choose exactly one of these values: INTERRUPTS_DIALOGUE / ALTERS_TONE_OR_MEANING / NO_CONVERSATIONAL_EFFECT. Judge the action in its full conversational context.

T6 system prompt

You are a friendly, warm, and conversational participant in an ongoing video call. Earlier conversation: {earlier} Most recent context: {recent} Continue the conversation naturally from this point in your own words. Keep watching and listening as the conversation unfolds. Respond naturally to anything the other person says or does without describing the video or these instructions. Start speaking now.

Qwen3.6-35B-A3B generates dialogue-history summaries from the pre-event transcripts. T6 inserts them into the system prompt; for T1, they serve only as the dialogue context given to the judge (Appendix B.1). Each summary has two parts: `earlier_history_summary` ($\leq$45 words, summarizing the earlier conversational background) and `recent_history_summary` ($\leq$85 words, describing the recent context before the takeover point). The summarization instruction requires referring to the replaced speaker as "you" and to the other participant as "the other person," and forbids revealing event labels, timestamps, or future content.

For T3 and T5, every adapter appends the post-prompt to the end of the system prompt with the prefix "At the instant the media stream ends, follow this final instruction:", and the model answers after playback ends. When the media stream ends, each adapter sends the API's native end-of-input signal; VITA-1.5 additionally receives the post-prompt again as a user turn.

T5 function-label definitions. The prompt attaches the following definitions to the three labels in the T5 post-prompt of Table 4:

T5 function-label definitions (attached to the T5 post-prompt)

- INTERRUPTS_DIALOGUE: The action asks the other participant to stop, pause, stay quiet, or yield the floor. It includes a clear bid to take the floor or an action that terminates ongoing speech.
- ALTERS_TONE_OR_MEANING: The dialogue continues, but the action communicates stance, emotion, or information that calls for the other participant to change what they say or how they say it. Examples include rejection, distress, disbelief, disapproval, relief, or a corrective cue.
- NO_CONVERSATIONAL_EFFECT: The movement only accompanies, illustrates, reenacts, or emphasizes the current speech. It does not ask the other participant to change their turn, content, or delivery.

B Scoring Details

Qwen3.8-Max (snapshot qwen3.8-max-0902, non-thinking mode, temperature = 0) performs all judging, and every model utterance the judge sees is the ASR transcript of the audio that actually played, except the T5 action description, which comes from the model's text answer. Boxes marked as excerpts omit output-format, anti-injection, and some auxiliary instructions; we release the full prompts with the code. The judges fall into two types. T1, T2, and T6 stop samples use a content-only judge (Appendices B.1 and B.2), and a program determines timing and sets a score cap (Appendix B.7). T4, T5, and the T6 adapt and continue samples use their own task judges (Appendices B.3–B.5). The T6 adapt and continue judge uses the scale in Table 5; the T4 judge uses its own rubric (Appendix B.3).

Table 5: Judge scale for the T6 adapt and continue samples.

Score	Meaning
0	No usable response, wrong direction, or target behavior absent
1	Role of the action largely missed, or response far too late
2	Correct direction but weak, generic, or awkward
3	Clear, timely, and appropriate to the conversational context
4	Specific and natural, and preserves the surrounding dialogue

B.1 T1 Content Judge

The judge receives only the dialogue context, the other person's utterance, and the evaluated model's transcript; it receives no event class, required floor act, timing outcome, latency, or deadline. Fol-

lowing the rubric below, the judge outputs a content score from 0 to 4 and quotes a span of the transcript as evidence. The program then takes the smaller of the content score and the timing cap as the final score.

Content-judge prompt (T1, T2, and T6 stop; excerpt)

Assess only the CONTENT of audible speech in a conversation. A separate program evaluates turn-taking and timing. You receive no event class, required floor act, timing verdict, latency, or numerical deadline. Do not infer these or grade them. ASR may duplicate or omit words. Ordinary hesitations, self-corrections, and incomplete clauses in spontaneous speech are not by themselves errors. Judge the recoverable meaning and its relationship to the supplied conversation context.

- 4 thread_preserved: specific, coherent content that preserves the conversational thread;
- 3 coherent: coherent content with a minor content defect or limited specificity;
- 2 rough: a specific content defect makes the meaning generic, repetitive or difficult to follow;
- 1 breakdown: a concrete major CONTENT contradiction supported by quoted words;
- 0 unusable: no intelligible content.

Never justify a score by when someone spoke, whether they stopped, or whether they should listen. The program combines your content score with its independently computed timing cap.

B.2 T2 AND T6 STOP CONTENT JUDGE

T2 and T6 stop samples use the same content-judge prompt as Appendix B.1 (T2 samples carry no dialogue context summary; Table 6), and the judge likewise receives no event class or timing information. Appendix B.7 gives the timing caps for both.

B.3 T4 RESPONSE JUDGE

First-response extraction. Starting at the expression onset, the evaluation takes the model's first audible response and extends it sentence by sentence (splitting at punctuation or at pauses of at least 0.8 s) until it accumulates 3 s of voiced speech, with at most 3 sentences and a span of at most 10 s; extension stops early at a gap longer than 1.5 s or when new user words are released. The judge sees the model's own speech from the 8 s before the expression only as context, and we do not score it.

The judge receives the expression class, the user's words spoken by the time the response starts, the selected first response (transcript, word-level timing, and first-voicing latency), the acoustic description of that segment, the user's words released during the response, and the original speaker's reaction and whether it changed after the expression as a reference. It outputs a score from 0 to 4 under the rubric below. The rubric assesses whether the response is appropriate in the situation at that time and does not require the model to change relative to an earlier vocal baseline; if the model has no audible response throughout the observation interval, the sample scores 0 directly.

T4 judge prompt (excerpt)

Score how appropriate the selected audible response is for the user's verified facial expression and the conversational context available at that time. This is an outcome score, not proof that vision caused the response or that the speaker changed tone from a previous baseline.
Evidence rules:

- User words are restricted to the input channel actually played to the agent. Words marked as available at response start may explain its opening. Words released during the response may explain only subsequent speech.
- The clip contains the first sentence or an ongoing remainder, with limited adjacent short units. Never assume how a cut-off sentence ends, or invent later speech. A continued sentence, including one resumed after a pause across the cue, is not evidence of rapid adaptation.
- Latency is first audible onset after client cue release, not visual processing/adaptation latency. Speech may have been generated before the cue and queued for playback.
- ASR is the spoken-content source. MOSS describes only directly audible delivery and vocalizations. Missing ASR words do not mean silence; laughter/sighs can be valid responses.
- No pre-cue acoustic baseline is supplied. Do not claim that pitch, volume or pace increased or decreased relative to earlier delivery.

- The human reaction is one possible reference, not a required script or required acoustic change. Do not require emotion naming, an explicit change, or copying the human. Do not cap a response merely because its appropriateness could also be explained by the audio.
- A short acknowledgment can earn full credit when it fits the moment. Length, verbosity and rich MOSS adjectives are not reasons to award more points.

Rubric (integer 0 to 4):

- 0: No audible response in the verified interval, or the response clearly conflicts with the user's emotion and conversation.
- 1: An audible response, but it largely misses the emotional or conversational needs of the moment.
- 2: Plausibly appropriate, but generic, awkward, insufficiently supported, or mismatched in intensity.
- 3: Clearly appropriate in content or supported delivery, and natural in the available context.
- 4: Clearly appropriate, specific to the available context, well timed for the conversational state, and preserves the dialogue. A fitting brief response can qualify.

The judge prompt also includes the following evidence rule. Before calling the judge, the program verifies that the selected window contains an audible response; samples without a response score 0 directly and skip the judge.

T4 evidence rule (excerpt)

The program has verified an audible response in the selected window. Do not reclassify it as silence. Selected transcript words are INCLUDED evidence, even when the aligner assigns the same start and end timestamp. A fragment caused by the evaluation boundary, or brevity alone, is not an independent defect. Zero requires evidence of a conflicting response; absence of a response is handled by the program before this API call. Copy a contiguous phrase from the selected transcript for ASR, or the supplied acoustic description for MOSS.

B.4 T5 Action-Description Judge

The judge receives the human reference action description and the model's action description, and outputs a score from 0 to 2 under the following rubric.

T5 action-description judge prompt (excerpt)

Judge whether the model captures the core visible movement or nonverbal behavior in the human reference. Ignore conversational function, intent, emotion, and consequences.
Apply a lenient semantic match to the visible action, using these rules consistently:

- Accept synonyms, paraphrases, ordinary action-family descriptions, minor spelling or word-spacing errors, and harmless differences in detail. Do not require the candidate to repeat the reference.
- Ignore left/right side, screen position, hand identity, spatial direction, repetition count, motion size, and omitted secondary details. Ignore added intent, emotion, conversational function, and consequences.
- Match only at the specificity supported by the reference. A broad reference may match either a broad paraphrase or a concrete, plausible instance of that action family. Missing detail in the reference is not evidence of a mismatch.
- If either description contains multiple actions, one shared distinctive core action earns 2. Harmless extra movements do not cancel that match. Explicitly denying the reference action does.
- A specific reference still requires meaningful action overlap: sharing only a body region or saying that it moves earns 1. A clearly different action earns 0, even if it uses the same body part.
- You have text descriptions, not the video. Do not invent unseen movements to create a match.

Rubric:

- 0: Missing visible-action content, an explicit contradiction, or clearly different actions with no meaningful action overlap.
- 1: Some action overlap, but only a body region or generic motion matches a specific reference.
- 2: The same core action, a reasonable action-family match (including a concrete instance of a broad reference), or a shared distinctive component of a compound action.

Calibration examples:

- Reference "stop gesture"; candidate "raises an open hand": 2.
- Reference "holds forehead and opens the other palm"; candidate "holds forehead": 2.
- Reference "illustrative gesture"; candidate "raises her index finger to emphasize a point": 2.
- Reference "nods"; candidate "moves the head": 1.

- Reference "stop gesture"; candidate "rubs the forehead": 0.
- Reference "nods"; candidate "shakes the head": 0.

The full prompt also contains further matching rules, anti-injection instructions, and 7 additional calibration examples, which we release with the code.

B.5 T6 Adapt and Continue Judge

T6 stop samples use the content judge of Appendix B.2, with the timing cap in Appendix B.7. Adapt and continue samples have no fixed deadline, and the following task judge scores them (the prompt calls these samples MODIFY and INVARIANT). The judge receives the event class, the verified human action description, the original speaker's reaction, the conversation prompts and source transcript around the action, the model transcript, the VAD speech intervals and latency metrics, and the acoustic description. The transcript and acoustic description both come from the same bounded evidence window (3 s before action onset to the end of observation), and the acoustic description splits into a pre-action baseline segment and a post-action response segment.

T6 adapt and continue judge prompt (excerpt)

Expected behaviors:

- MODIFY: The action changes the other person's stance or delivery. A strong response changes its content, attitude, or vocal delivery in a way that fits the context.
- INVARIANT: The action is conversationally incidental. A strong response remains coherent and natural without an unwarranted pause, topic shift, or emotional change.

For MODIFY and INVARIANT, objective speech intervals are authoritative. There is no fixed latency deadline for MODIFY or INVARIANT. For INVARIANT, judge continuity in context: ordinary listening, sentence-final pauses and brief responses are allowed. Do not require a minimum fraction of speech before or during the action. Low pre-event activity is neither a success nor a failure by itself. No usable response in the observation window scores 0. A generic response compatible only with the spoken words does not demonstrate a MODIFY response.

Use this common scale, interpreted for the expected behavior:

- 0: no usable response, the direction is wrong, or the target behavior does not occur;
- 1: the action's role is largely missed, or the response is far too late;
- 2: the response moves in the correct direction but is weak, generic, or awkward;
- 3: the response is clear, timely, and appropriate to the conversational context;
- 4: the response is specific and natural and preserves the surrounding dialogue.

For MODIFY, score 0 for a contradictory change, 1 for responding only to the words with no action related change, 2 for a weak or excessive adjustment, 3 for a clear fitting adjustment, and 4 for a specific contextual adjustment. For INVARIANT, score 0 for a false stop or derailment, 1 for a major unwarranted shift, 2 for a minor disruption, 3 for natural continuation, and 4 for complete thread preservation.

B.6 Judge Input Fields

Table 6 lists the input fields that each task's judge model actually receives. All fields pass in JSON format with temperature = 0. The content judge (T1, T2, and T6 stop) returns `quality_score`, `content_assessment`, `evidence_quote`, and `content_reason`; the T4 judge returns `score`, `assessment`, `evidence_source`, `evidence_quote`, and `reason`; the T5 and T6 judges return `score` and `reason`.

The T1 and T2 columns and the T6 stop samples use the content-only judge, and a program determines timing separately. The T6 column reflects the adapt and continue samples. The T5 judge compares only the model's and the human's action descriptions and involves no dialogue context or timing metrics.

Table 6: Input fields provided to the judge in each task (✓ = provided, – = not provided).

Input field	Source	T1	T2	T4	T5	T6
Event class (gold class)	Annotation	–	–	✓	–	✓
Dialogue context summary	Qwen3.6-35B-A3B	✓	–	–	–	–
Conversation prompts and source transcript	Source corpus	–	–	–	–	✓
Other person's utterance	Source-corpus transcript	✓	✓	✓	–	✓
Model response transcript	Qwen3-ASR	✓	✓	✓	–	✓
Model pre-event speech (context only)	Qwen3-ASR	–	–	✓	–	–
Speech intervals and latency metrics	VAD	–	–	✓	–	✓
Acoustic description	MOSS-Audio	–	–	✓	–	✓
Human reference response	Annotation	–	–	✓	–	✓
Human action description	Annotation	–	–	–	✓	✓
Model action description	Model output	–	–	–	✓	–

B.7 Formal Definition of Timing Caps

For T1, T2, and T6 stop samples, r is the content score from the content judge (Appendices B.1 and B.2), and the program determines the final score r' by the rules below, which is equivalent to taking the smaller of r and the timing cap. Within each task, the program checks conditions from top to bottom and applies the first match. "Stops speaking" means a sustained silence of at least 0.8 s (Appendix E.2); δ is the latency from event onset to stopping speaking, t_{onset} is the onset time of the model's new speech, and t_{ref} is the reference response time. The yield deadline $\tau = 1.5$ s is the same for T1 and T6 and follows the upper bound on human repair time reported by Full-Duplex-Bench v1.5 (Lin et al., 2026b).

$$r'_{\text{T1}} = \begin{cases} 0 & \text{not speaking at } t_{\text{event}} \\ 0 & \text{interruption, model does not stop speaking} \\ \min(r, 1) & \text{interruption, } \delta > \tau \\ 0 & \text{backchannel, } \delta \leq \tau \text{ and no resumed speech afterward} \\ \min(r, 1) & \text{backchannel, } \delta \leq \tau \text{ but speech resumes afterward} \\ r & \text{otherwise} \end{cases} \tag{2}$$

$$r'_{\text{T2}} = \begin{cases} 0 & \text{turn yield, model already speaking at the event time} \\ 0 & \text{turn yield, model never starts speaking} \\ \min(r, 1) & \text{turn yield, } t_{\text{onset}} > t_{\text{ref}} + 0.5\,\text{s} \\ \min(r, 1) & \text{pause, model speaks during the pause} \\ 4 & \text{pause, model stays silent (no judge call)} \\ r & \text{otherwise} \end{cases} \tag{3}$$

$$r'_{\text{T6-stop}} = \begin{cases} 0 & \text{not speaking at } t_{\text{event}} \\ 0 & \text{model does not stop speaking} \\ \min(r, 1) & \delta > \tau \\ r & \text{otherwise} \end{cases} \tag{4}$$

$$r'_{\text{T6-cont}} = \begin{cases} 0 & \text{no model speech from action onset to the end of observation} \\ r & \text{otherwise} \end{cases} \tag{5}$$

T6 continue samples have no speaking-ratio threshold: as long as the model produces a usable response after action onset, the judge decides semantically whether the response continues the conversation naturally. Ordinary listening, sentence-final pauses, and brief responses incur no penalty from the timing rules.

B.8 JUDGE VALIDATION

We validate the Qwen3.8-Max judge in two ways: we compare its scores with those of two judge models from other vendors, GPT-5.6-sol and GLM-5.3-Flash, on 200 items, and we manually review the scoring results. We sample from completed judge calls, stratifying by task and the original Qwen score to cover each score level. The sample contains 35 T1, 30 T2, 50 T4, 30 T5, and 55 T6 items. T4 includes 40 audiovisual and 10 audio-only items; T6 includes 15 stop items and 40 adapt or continue items, including 10 audio-only items. Sampling balances the seven evaluated models where possible; the T1 subset also includes four original-speaker recordings.

GPT and GLM rescore the items with the same rubrics and the same inputs as the original judge, including transcripts, speech intervals, and acoustic descriptions where applicable, and neither sees the original scores. Three independent GPT-5.6-sol sessions (medium reasoning effort) each score all 200 items, and we take the per-item median as the GPT score. GLM-5.3-Flash scores one item per request with temperature 0.

Table 7 reports quadratic-weighted Cohen's κ for each model pair. We compute 95% percentile bootstrap intervals using 2,000 item resamples within each reported group, with seed 20260926. We report the 30 T5 action-description scores on their 0–2 scale separately from the 170 items on the 0–4 scale. Because the program computes timing separately, the comparison uses raw judge scores before timing caps.

Table 7: Agreement across three judge models. Cells give quadratic-weighted κ with 95% bootstrap intervals. GPT denotes the median of three GPT-5.6-sol sessions.

Rubric group	Items	Qwen/GPT	Qwen/GLM	GPT/GLM
Content (T1, T2, T6 stop)	80	0.845 [0.760, 0.909]	0.744 [0.603, 0.841]	0.693 [0.552, 0.798]
T4	50	0.861 [0.752, 0.934]	0.771 [0.660, 0.845]	0.736 [0.581, 0.844]
T6 adapt and continue	40	0.798 [0.629, 0.906]	0.850 [0.732, 0.921]	0.816 [0.678, 0.905]
T5 action description (0–2)	30	0.727 [0.479, 0.897]	0.791 [0.624, 0.904]	0.933 [0.837, 1.000]
All 0–4 scores	170	0.837 [0.775, 0.886]	0.779 [0.704, 0.838]	0.740 [0.659, 0.807]

On the 170 items scored from 0 to 4, Qwen agrees within one point with GPT on 91.8% of items (95% CI: 87.6–95.9%) and with GLM on 92.4% (88.2–95.9%), with Spearman correlations of 0.847 (0.787–0.897) and 0.793 (0.717–0.852). The three judges together reach an ordinal Krippendorff's α of 0.781 (0.719–0.830). Because judges from three vendors score independently and still agree closely, the scores reflect the rubrics rather than the preferences of a single model. The authors also manually reviewed the scoring results and found the judge scores appropriate under the rubrics.

C ACOUSTIC ANALYSIS PROMPTS

T4 and T6 use MOSS-Audio-4B-Instruct to extract acoustic descriptions from model-generated audio. The two tasks use different prompts to match their analysis scopes.

T4 acoustic prompt. The input is the first-response audio selected by the rules in Appendix B.3.

T4 acoustic prompt (MOSS-Audio-4B-Instruct)

Listen only to this selected first-response audio clip. It may start in the middle of a sentence and may include adjacent short sentences and their pauses. Describe directly audible content, tone, pitch, energy, pace, pauses and vocalizations such as laughter, sighs or gasps. Do not infer any unseen facial expression or assign a score. Do not infer vocal emotion solely from the meaning of words. Do not infer age, personality, intentions, or change from a previous voice baseline that was not provided. Use ASR separately for exact words; do not invent missing words. Say unclear if the clip is too short or ambiguous. Keep it concise.

T6 acoustic prompt. The program splits the evidence window (3 s before action onset to the end of observation) at action onset into a baseline segment and a response segment, and describes each segment with the same prompt:

T6 acoustic prompt (MOSS-Audio-4B-Instruct)

Describe the audible voice, including its pitch, energy, speaking pace, vocal tone, pauses, and non-speech vocalizations. Focus on acoustic characteristics. Keep the description concise.

Both tasks use greedy decoding (temperature = 0) with at most 384 generated tokens for T4 and 256 for T6. The prompts ask only for audible acoustic evidence and no quality judgment; ASR supplies the word content separately, and the acoustic description then serves as input to the judge model.

D Models and Experimental Setup

D.1 Access, Input, and Evaluation Scale

Model access. We call the proprietary and lightweight models through each vendor's real-time API (the Google GenAI SDK for Gemini and WebSocket for the others). MiniCPM-o 4.5 supports both a cloud API and local deployment. We deploy VITA-1.5 locally on four NVIDIA H800 80 GB GPUs with two shards.

Unified input protocol. We extract video as JPEG frames at 1 fps (720 px longest side, ffmpeg quality `-q:v 5`) and resample audio to 16 kHz mono PCM. Media arrives in one-second units in the original temporal order: at each one-second boundary, the adapter sends that second's audio chunk together with the corresponding video frame. The audio-only condition keeps the full audio track and removes only the video frames.

Inference settings. We use each API's default temperature, except for the two Qwen real-time models, which we set to 0.7. For the recognition tasks (T3, T5), we disable the model's built-in voice activity detection and trigger the model's answer after media playback ends.

Evaluation scale. Each sample runs once per condition (run_id = 0, seed = 0). T1, T2, T3, and T5 run only under the audiovisual condition; T4 and T6 run under both audiovisual and audio-only conditions, with the order counterbalanced by a hash of the sample ID. Each model runs 874 independent sessions in total, about 9.5 hours of streamed media.

Judge and tool models. The judge model is Qwen3.8-Max (Qwen Team, 2026), which we call with the fixed snapshot qwen3.8-max-0902 in non-thinking mode through an OpenAI-compatible API with temperature = 0. Qwen3-ASR-1.7B transcribes the output audio, and Qwen3-ForcedAligner-0.6B aligns word-level timestamps. MOSS-Audio-4B-Instruct (Yang et al., 2026a) extracts the acoustic descriptions for T4 and T6. Qwen3.6-35B-A3B generates the dialogue-history summaries used in the T6 prompts and in the T1 judge context.

D.2 Adapter Parameters

Table 8 lists the key adapter parameters that differ across models. For spoken-behavior tasks, the adapter streams 3 s of trailing silence under the audio-only condition (none under the audiovisual condition) and then waits 4 s for the response; for recognition tasks, it streams 8 s of trailing silence and waits up to 12 s for the answer. We resample all output audio to 24 kHz PCM16.

GPT-Realtime-2.1 requires 24 kHz input, so the adapter resamples from the standard 16 kHz pipeline in real time. MiniCPM-o 4.5 requires float32 format and 1000 ms audio chunks, and the adapter performs the corresponding conversion. VITA-1.5 has no off-the-shelf WebSocket adapter, so we call it through a custom interface on the local deployment.

Table 8: Adapter parameters that differ across models (– = not configurable or not applicable).

Model	Transport	Input audio	Chunk	Voice	VAD mode	Temp.
Gemini 3.8 Live	GenAI SDK	16 kHz PCM16	40 ms	Kore	Native (auto)	Default
Gemini 3.1 Flash Live Preview	GenAI SDK	16 kHz PCM16	40 ms	Kore	Native (auto)	Default
GPT-Realtime-2.1	WebSocket	24 kHz PCM16	40 ms	Marin	Semantic VAD	Default
Qwen3.5-Omni-Plus-Realtime	WebSocket	16 kHz PCM16	40 ms	Tina	Semantic VAD	0.7
Qwen-Omni-Turbo-Realtime	WebSocket	16 kHz PCM16	40 ms	Tina	Server VAD	0.7
MiniCPM-o 4.5	WebSocket	16 kHz float32	1000 ms	–	Native (duplex)	Default
VITA-1.5	Local	16 kHz	–	–	Native (duplex)	–

E Clip Construction and VAD Parameters

E.1 Clip Duration

Each task's clips are cut around the annotated event and keep context and continuation of different lengths. Clips for the turn-level tasks (T1, T2) have median durations of 35–40 s; expression-recognition clips are fixed at 21 s; response clips have median durations of 34–51 s.

T1 (Interruption Handling): The event occurs at about 20 s. The clip extends to the later of 20 s after the event or the end of the other person's complete turn, but not beyond the end of the original recording.

T2 (Turn Entry): The decision point lies at 15 s, which provides 15 s of natural dialogue context. HOLD clips keep the real pause and at least 5 s after the speaker resumes. TAKE clips keep the other person's audio and video after the turn yield. The clip ends at the later of 20 s after the decision point or the end of the next complete turn.

T3 (Expression Recognition): 20 s of dialogue context plus the 1 s target expression at the end; the clip ends when the expression ends and lasts a fixed 21 s.

T4 (Expression-Conditioned Response): 20 s of context plus the 1 s expression event, followed by up to 30 s to observe the model's response; the median duration is 51 s.

T5 (Action-Function Recognition): Up to 20 s of context, ending when the annotated action ends. In 59 samples the action lies at 20 s; the 9 shorter clips have 2.7–19.4 s of context.

T6 (Action-Conditioned Response): Up to 20 s of context, plus the complete annotated action, plus 10 s of the original conversation after the action.

E.2 Voice Activity Detection Parameters

T1, T2, and T6 share the same VAD. It first applies gain normalization to the output audio, then uses Silero VAD to estimate the per-frame speech probability, and finally refines speech-segment boundaries with an energy threshold. Gain normalization changes no timestamps. Table 9 lists the parameters.

Table 9: Voice activity detection parameters.

Parameter	Value
Speech probability threshold (enter / exit)	0.5 / 0.35
Minimum silence	200 ms
Minimum speech	100 ms
Boundary-refinement energy threshold	−45.0 dBFS
Boundary-refinement frame length	20 ms
Sustained silence for yield detection (T1, T6)	800 ms

T1 and T6 stop treat a sustained silence of at least 800 ms as stopping speaking. T4 first-response extraction uses the speech intervals recorded at generation time (Appendix B.3).

F Annotation Procedures

F.1 Candidate Event Extraction

The benchmark automatically extracts three kinds of candidate events from the source corpus's pre-computed metadata. Turn events (interruptions, backchannels, within-turn pauses, and turn yields) come from voice activity segments and word-level transcripts, identified by speech-overlap patterns and pause-duration rules. Expression events come from per-frame emotion-classification logits, located by softmax probability thresholds and consecutive-frame detection. Action events come from 30 Hz head, body, and hand kinematic trajectories, detected by joint motion amplitude and direction changes. All candidate events must also satisfy context constraints: sufficient dialogue context before the event and observable subsequent interaction after it. Expression annotation additionally asks annotators to judge whether the original speaker's response changed because of the expression; action annotation asks annotators to determine the action's functional effect on the dialogue.

F.2 Class Distribution

Table 10 lists the class distribution within each event group. Differences in class counts come from how often each event actually occurs in natural conversation.

Table 10: Class distribution within each event group.

Event group	**Class**	**Count**
Interruption / backchannel	Interruption	130
	Backchannel	49
Pause / turn yield	Within-turn pause	94
	Turn yield	100
Facial expression	Happiness	22
	Surprise or fear	21
	Sadness	20
	Disgust	19
	Anger	12
	Contempt	5
Action function	No conversational effect	26
	Interrupts dialogue	21
	Alters tone or meaning	21

F.3 Annotation Quality Control

Three independent annotators review each candidate event, and a majority vote decides the final label. When no two annotators agree, a fourth senior annotator arbitrates. Each new annotator completes 10 practice samples before formal annotation.

Expression annotation (T3/T4): annotators view the target person's video together with both speakers' synchronized audio, and confirm or correct the automatically extracted expression class and temporal boundaries. Disagreements on temporal boundaries larger than 0.5 s trigger arbitration.

Turn annotation (T2): annotators view speaker-separated dual-channel audio with subtitles and judge whether the target speaker has yielded the floor at the annotated time.

Action annotation (T5/T6): annotators watch actor-side video trimmed exactly to the start and end of the annotated action and provide open-ended action descriptions.

F.4 T5 Human Action References

Each of the 68 T5 samples has a short action description written by one annotator, which the judge model uses to verify the model's action description. The reference descriptions cover more than

20 action types, including stop gestures, nods, head shakes, waves, holding the forehead, mimed actions, and clapping.

G HUMAN REFERENCE ON T1

We evaluate the original speaker's recordings on all 179 T1 events and obtain a human reference score of **82.17/100**. The recording intervals match the corresponding model samples and include 130 genuine interruption events (P) and 49 backchannel events (N1). These recordings reflect natural behavior in the original conversations.

The human recordings go through exactly the same evaluation pipeline as the models: VAD (Appendix E.2), Qwen3-ASR-1.7B transcription, Qwen3-ForcedAligner-0.6B alignment, Qwen3.8-Max content judging (Appendix B.1), and the T1 timing cap (Appendix B.7). The final score is 25 times the mean of the two class-average ratings, and all samples remain in their class denominators. The human clock uses the original recording time and excludes the latency of model input transport and the playback chain. We keep the per-item audio hashes, aligned transcripts, judge outputs, and score records.

Table 11: T1 scores and timing behavior of the original speaker's recordings.

Event class	Samples	Timing correct	Score /100
Genuine interruption (P)	130	104	74.04
Backchannel (N1)	49	49	90.31
All events	179	153	**82.17**

The score over all events is a class-balanced average, not a sample-weighted one. For P, correct timing requires speaking at the event and stopping within 1.5 s; for N1, it requires speaking at the event and not entering, within 1.5 s, a stop that meets the sustained-silence condition. The timing accuracy is $(104/130 + 49/49)/2 = \mathbf{90.00}\%$.

Of the 130 P recordings, 129 show a measurable stop: 104 stop within 1.5 s and 25 stop late.

Table 12: Reaction latencies of the original speaker for stopping (T1) and starting to speak (T2).

Measure	Measurable / total	Mean (s)	Median (s)	P90 (s)
T1 P: stop before VAD sustained silence	129/130	**1.238**	**0.400**	**2.976**
T2 TAKE: annotated turn-taking onset	100/100	**1.072**	**0.900**	**2.026**

For T1, we measure from interruption onset to the start of the first silence interval lasting at least 0.8 s within the observation window, and the statistics include all 25 late samples. For T2, we use the annotations of all 100 TAKE events and subtract the annotated turn-yield time from the original speaker's speech onset. P90 denotes the 90th percentile. Stopping and starting measure different actions, so we do not merge them into one mean latency.

H PER-CLASS RESULTS

Tables 13–16 give per-class scores (0–100) for each task under the audiovisual condition. For T1, T2, T4, and T6, each class score is 25 times the class-average rating; for T3 and T5, it is the per-class F1. The last row averages the seven models.

T2 averages 74.2 on the pause class but only 24.4 on the turn-yield class: models tend to keep waiting, and failures occur mainly at turn yields. The T1 interruption and backchannel classes score similarly (18.9 and 19.4).

Every model has an F1 of 0 on the contempt class (which has only 5 samples). Happiness is the easiest class to recognize (average 59.3); apart from contempt, negative expressions such as disgust and anger are the hardest (13.0 and 18.3).

Table 13: Per-class scores on the Interaction Capacity tasks (0–100).

Model	T1 Interruption	T1 Backchannel	T2 Pause	T2 Turn yield
Qwen3.5-Omni-Plus-Realtime	29.4	24.5	39.6	39.8
Gemini 3.1 Flash Live Preview	22.5	24.0	94.9	27.0
Gemini 3.8 Live	21.5	12.8	89.6	27.5
GPT-Realtime-2.1	13.7	7.1	98.9	16.5
MiniCPM-o 4.5	17.3	41.8	90.4	19.5
Qwen-Omni-Turbo-Realtime	12.1	5.1	76.7	22.0
VITA-1.5	15.6	20.4	29.0	18.5
Model average	*18.9*	*19.4*	*74.2*	*24.4*

Table 14: Per-class scores on T3 expression recognition (F1, 0–100).

Model	Anger	Contempt	Disgust	Happiness	Sadness	Surprise or fear
Qwen3.5-Omni-Plus-Realtime	35.3	0.0	9.1	56.0	58.3	40.0
Gemini 3.1 Flash Live Preview	12.5	0.0	10.0	60.3	49.1	33.3
Gemini 3.8 Live	47.1	0.0	0.0	50.0	43.2	38.7
GPT-Realtime-2.1	33.3	0.0	27.3	61.5	56.0	35.3
MiniCPM-o 4.5	0.0	0.0	0.0	54.5	46.4	17.4
Qwen-Omni-Turbo-Realtime	0.0	0.0	9.1	75.0	60.4	57.1
VITA-1.5	0.0	0.0	35.3	57.7	19.4	34.8
Model average	*18.3*	*0.0*	*13.0*	*59.3*	*47.5*	*36.7*

T4 scores vary less across classes than T3 scores: even though the T3 F1 on contempt is 0, model responses to contempt still average 57.9. This is consistent with Finding 3: recognizing a cue and acting on it are largely decoupled.

Interrupting actions are the hardest to both recognize and respond to: the T5 F1 on the interrupts-dialogue class averages only 20.9, and T6 stop samples average only 17.3, below adapt (36.7) and continue (40.7) samples.

Table 15: Per-class scores on T4 expression-conditioned response (0–100).

Model	Anger	Contempt	Disgust	Happiness	Sadness	Surprise or fear
Qwen3.5-Omni-Plus-Realtime	66.7	80.0	55.3	80.7	66.2	79.8
Gemini 3.1 Flash Live Preview	62.5	80.0	51.3	67.0	46.2	58.3
Gemini 3.8 Live	64.6	60.0	48.7	67.0	50.0	64.3
GPT-Realtime-2.1	35.4	40.0	21.1	40.9	56.2	57.1
MiniCPM-o 4.5	37.5	55.0	44.7	52.3	28.7	51.2
Qwen-Omni-Turbo-Realtime	37.5	45.0	32.9	52.3	33.8	41.7
VITA-1.5	41.7	45.0	44.7	51.1	47.5	56.0
Model average	*49.4*	*57.9*	*42.7*	*58.8*	*47.0*	*58.3*

Table 16: Per-class scores on the action tasks (grounded F1 for T5, judge scores for T6; 0–100).

	T5			T6		
Model	Interrupts dialogue	Alters tone or meaning	No conv. effect	Stop	Adapt	Continue
Qwen3.5-Omni-Plus-Realtime	24.0	50.0	69.6	33.3	64.3	57.7
Gemini 3.1 Flash Live Preview	48.3	32.3	53.7	11.9	47.6	43.3
Gemini 3.8 Live	37.0	22.2	39.0	20.2	35.7	37.5
GPT-Realtime-2.1	37.0	36.4	24.2	6.0	20.2	18.3
MiniCPM-o 4.5	0.0	25.0	7.4	10.7	32.1	45.2
Qwen-Omni-Turbo-Realtime	0.0	10.8	0.0	16.7	25.0	40.4
VITA-1.5	0.0	17.1	0.0	22.6	32.1	42.3
Model average	*20.9*	*27.7*	*27.7*	*17.3*	*36.7*	*40.7*

I Limitations and Future Work

First, all samples come from English two-person web video calls in Seamless Interaction and do not cover face-to-face exchanges, multi-party conversations, or other languages. Second, expressions cover only six classes and actions distinguish only three conversational functions; the benchmark does not yet include other nonverbal cues such as gaze, head pose, and laughter. Third, each sample centers on a single event, and the benchmark evaluates only the model's speech output; it does not evaluate how cues accumulate over long multi-turn conversations, nor does it assess model-generated video or avatar behavior. Fourth, the results for proprietary models reflect the API versions we accessed in September 2026. Future work can extend the benchmark to multilingual, multi-party, and face-to-face settings and include more types of nonverbal cues.

J Case Studies

We select four examples from the final evaluation records to illustrate the failures reported in Section 4. Model utterances are ASR transcripts with punctuation added; | marks the event onset. Times are relative to event onset.

Case 1: Late stop after an interruption (T1_0078, GPT-Realtime-2.1)

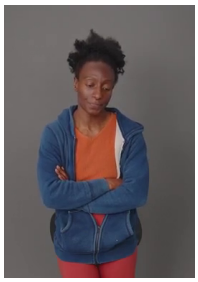
Context. The model continues the replaced speaker's turn about a fitness challenge. The user interrupts: "Can you imagine what the athletes go through physically by experiencing the endorphin high".

Model. "... or just enjoying the identity | you've earned. You can still have momentum without chasing a finish line every time."

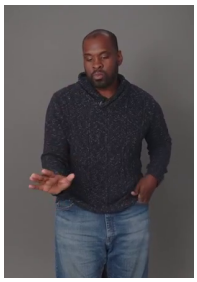
Timing. Speaking at interruption onset; stops after 4.54 s, beyond the 1.5 s limit.

Score. Content 3, capped to 1.

Original speaker. Stops after 0.40 s and answers with "Yeah. Yeah."; final score 3.



(a) Case 3: SADNESS



(b) Case 4: stop gesture

Figure 4: **Keyframes of the user video in Cases 3 and 4.** Each frame is taken at the event.

Case 2: Missed and late turn entry (T2_0029)

Context. The user finishes a question and yields the turn: "I have a question for you. Can you recall your happiest childhood memory?" The original speaker starts answering after 0.61 s, which sets the deadline at 1.11 s.

GPT-Realtime-2.1. Stays silent for the full 20 s observation window; final score 0.

MiniCPM-o 4.5. Starts after 2.37 s: "Oh, absolutely, I can totally picture it now. It was one of those perfect summer days…" Content 4, capped to 1.

Case 3: Recognized expression, inappropriate response (T4_REACTION_0095)

Context. The user talks about a relative near the end of life: "…Realistically, he doesn't have a lot of time left. Like, it should be a hospice situation. What's the point? What is the point?" Annotated expression: SADNESS (Figure 4a).

Gemini 3.1 Flash Live Preview, T3. Outputs SADNESS (correct).

Gemini 3.1 Flash Live Preview, T4 audiovisual. "It looks like you're getting pushed to one side. There, are you all right?" Score 0; the judge notes that the response ignores the user's grief and comments on the user's body movement instead.

Gemini 3.1 Flash Live Preview, T4 audio-only. "…Do you feel like his situation is starting to affect you?" Score 3.

Qwen3.5-Omni-Plus-Realtime, T4 audiovisual (T3: SADNESS). "I hear how much that hurts. When things feel this heavy, it's normal to wonder what any of it means." Score 4; the acoustic description reports a soft, slow, and soothing delivery.

Case 4: Described gesture, missed stop (T6_REACTION_0001, Gemini 3.1 Flash Live Preview)

Context. The user asks the model to help more with household chores. While the model speaks, the user extends a palm to signal stop and says "I mean, let's just, I'm just, look, I know you're doing…" Annotated function: INTERRUPTS_DIALOGUE, a stop sample (Figure 4b).

T5. "ACTION: pushing down palms; LABEL: ALTERS_TONE_OR_MEANING." The action description scores 2 of 2, but the function label is wrong.

T6 audiovisual. "Oh, I see. | You're feeling like I'm not pulling my weight around the house lately. Can you give me…" Stops after 3.98 s; content 3, capped to 1.

T6 audio-only. "Oh, I understand. | Yeah." Stops after 0.66 s; final score 3.

Cases 1 and 2 illustrate the two timing failures behind Finding 2. In Case 1, the model keeps talking for 4.5 s after the interruption, whereas the original speaker yields within half a second. In Case 2, one model never takes the turn after a direct question, and another gives a specific, fluent answer that arrives too late, so a content score of 4 becomes a final score of 1. Cases 3 and 4 illustrate the perception-to-interaction gap in Finding 3. In Case 3, the model labels the user's sadness correctly, yet its audiovisual response ignores the grief and reacts to incidental body movement, while Qwen3.5-Omni-Plus-Realtime acknowledges the grief in both wording and delivery. In Case 4, the model describes the stop gesture correctly but reads it as a change in tone, and keeps speaking for almost 4 s. In both cases the audio-only response scores higher, consistent with Finding 4.