TIMEACTBENCH: BENCHMARKING TIME-AWARE ACTIONS IN REAL-TIME INTERACTIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal precision is fundamental to human-level intelligence, allowing actions to be coordinated with events as they unfold. Recent real-time interactive models enable continuous perception and timely responses during ongoing interactions, opening the way to evaluating this capability beyond traditional turn-based settings. However, existing temporal benchmarks primarily emphasize retrospective temporal understanding or specialized forms of timing-sensitive interaction, leaving the ability to determine *whether and when to execute task-specific actions* under causal streaming observations insufficiently characterized. To address this gap, we introduce **TimeActBench**, a benchmark for systematically evaluating time-aware actions in real-time interactive models. TimeActBench comprises seven task families and 196 evaluation instances spanning audio, video, and Audio-Visual settings, and organizes time-aware action into two complementary dimensions: Perception and Detection (PD), which evaluates whether models can detect task-relevant events and emit the prescribed detection action when those events become observable, and Response and Interaction (RI), which evaluates whether models can execute appropriate actions according to temporal constraints imposed by an evolving interaction. Our main comparison evaluates 15 proprietary and open-source real-time interactive models, and we conduct a human study under the same evaluation protocol. Our results reveal substantial limitations in current models, with a pronounced overall gap between model and human performance across action correctness and temporal precision. These findings establish time-aware action as a challenging and underexplored capability of real-time interactive models. We expect TimeActBench to provide a systematic testbed for studying this capability and to advance research on temporal intelligence in real-time interactive models, and it will be fully open-sourced.

1 INTRODUCTION

Temporal precision is fundamental to human-level intelligence, enabling people to coordinate their actions with a changing environment. Even a correct action can fail when mistimed, as an early response may interrupt a speaker and a delayed control command may miss a moving target. Traditional turn-based models receive a completed input before producing a response, leaving little scope to decide when to intervene as events unfold. Recent real-time interactive models overcome this interaction constraint by processing continuous streams and producing outputs while new observations arrive (Défossez et al., 2024; Wang et al., 2025a; Cui et al., 2026; Thinking Machines Lab, 2026; OpenAI, 2026a; Ouyang & Jaganathan, 2026; Alibaba, 2026; ByteDance, 2026a;b). This opens the way to *time-aware action*, the ability to determine *whether an action is warranted and when to execute it* during an ongoing interaction. We use *action* to encompass observable linguistic responses, detection signals, and control operations.

Existing offline temporal benchmarks assess capabilities such as event ordering, duration estimation, event localization, and motion understanding (Liu et al., 2024; Cai et al., 2024). These tasks, requiring retrospective temporal understanding, measure how well models interpret temporal information in an observed sequence. Acting during an ongoing stream imposes an additional requirement: future observations are unavailable, and the model must decide whether the evidence available now is sufficient to trigger an action. Notably, retrospective understanding is not confined to offline benchmarks: streaming benchmarks such as SVBench, OVO-Bench, and RIVER likewise pose retrospective questions about the content observed up to the current moment (Yang et al., 2025; Niu

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

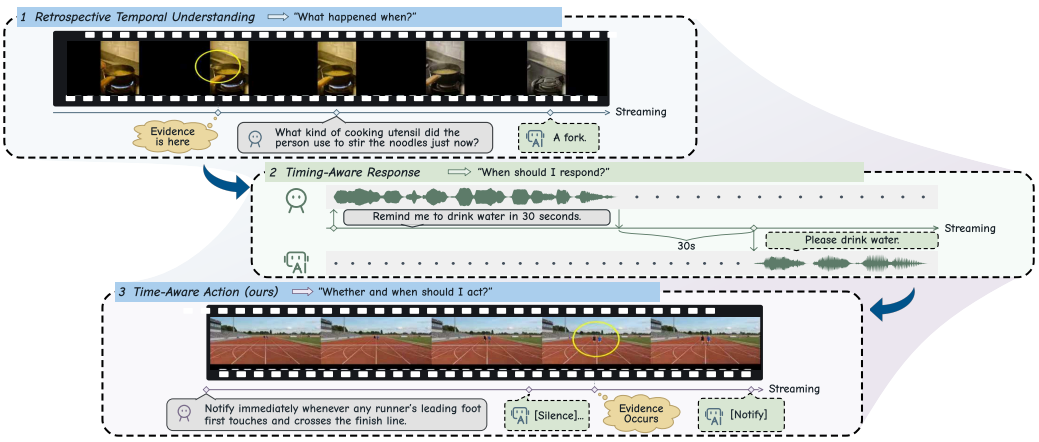


Figure 1: Three tiers of online streaming temporal reasoning.

Table 1: Comparison of temporal and interactive benchmarks. Causal streaming excludes future observations; timing evaluated means output timing contributes to evaluation. Control actions include task-level operations such as key presses and start–stop commands, implemented through tokens or tool calls. A/V/AV denote Audio/Video/Audio-Visual.

Benchmark	Modality	Causal Streaming	Timing Evaluated	Control Actions	Evaluated Output	Temporal Focus
TempCompass	V	✗	✗	✗	QA / Caption	Temporal Understanding
TemporalBench	V	✗	✗	✗	QA / Caption	Temporal Understanding
StreamingBench	AV	✓	✗	✗	QA	Streaming Understanding
SVBench	V	✓	✗	✗	Multi-turn QA	Temporal Dialogue
OVO-Bench	V	✓	✓	✗	QA / Response	Streaming Understanding
ProactiveVideoQA	V	✓	✓	✗	Response	Proactive Response
StreamReady	V	✓	✓	✗	Answer	Answer Readiness
RIVER	V	✓	✓	✗	Response	Online Interaction
Full-Duplex-Bench	A	✓	✓	✗	Dialogue Behavior	Turn Coordination
Game-Time	A	✓	✓	✗	Spoken Behavior	Temporal Coordination
VideoFDB	AV	✓	✓	✗	Conversational Behavior	AV Interaction
OmniInteract	AV	✓	✓	✗	Response	Streaming Interaction
TimeActBench (Ours)	**A / V / AV**	✓	✓	✓	**Task-Specific Action**	**Time-Aware Action**

et al., 2025; Shi et al., 2026). A model that can retrospectively explain *what happened when* may still fail to act at the appropriate moment.

Online benchmarks evaluate answer readiness, proactive responses, and conversational coordination under streaming observations (Lin et al., 2026b; Wang et al., 2025b; Shi et al., 2026; Lin et al., 2025; Lu et al., 2026). A common evaluation of task-specific detection signals, linguistic responses, and control operations across modalities remains underexplored. Figure 1 illustrates representative evaluation formats along this progression; Table 1 compares representative benchmarks, with further discussion in Sec. 2.

To this end, we introduce TimeActBench, which evaluates prescribed actions under causal streaming observations: models must act when a task's temporal condition is satisfied and remain inactive otherwise. Its seven task families and 196 instances span audio, video, and Audio-Visual streams. *Perception and Detection (PD)* covers event-triggered detection, while *Response and Interaction (RI)* covers actions governed by delays, synchronization, state transitions, and interaction signals. Temporally controlled inputs and fine-grained reference timestamps enable joint assessment of action correctness and temporal precision.

We evaluate proprietary and open-source real-time interactive models on their supported modalities and compare them with humans under the same causal streaming protocol. Across all 196 instances, the highest model F1 is 33.7%, compared with 95.0% for humans. The joint evaluation also exposes discrepancies between correctness and timing: low timing error among matched actions can coexist with frequent false activations. Increasing reasoning effort does not consistently improve performance and can increase timing error. These findings identify reliable action selection and timely execution as complementary challenges for real-time interactive models.

Our contributions are threefold:

- We formulate **time-aware action** as an evaluation problem that measures whether and when real-time models execute prescribed actions under causal streaming observations.
- We introduce TimeActBench, with seven task families and 196 instances spanning audio, video, and Audio-Visual inputs. Its PD and RI dimensions, temporally controlled inputs, and fine-grained annotations support joint assessment of action correctness and temporal precision.
- We establish model and human baselines, reveal a substantial overall human–model gap, and show that low matched-action timing error and increased reasoning effort do not by themselves establish reliable time-aware action.

2 Related Work

2.1 Real-Time Interactive Models

Moshi and Freeze-Omni support continuous spoken interaction through full-duplex or simultaneous listening and speaking (Défossez et al., 2024; Wang et al., 2025a). Proprietary systems include GPT Live and GPT Realtime, Gemini Live, Grok Voice, Qwen Realtime, SeedDuplex, and SeedRealtime (OpenAI, 2026a;b; Ouyang & Jaganathan, 2026; xAI, 2026; Alibaba, 2026; ByteDance, 2026a;b). Beyond speech, Thinking Machines Lab studies time-aligned multimodal micro-turns, and MiniCPM-o 4.5 supports full-duplex omni-modal interaction (Thinking Machines Lab, 2026; Cui et al., 2026). JoyAI VL Interaction and MOSS VL process continuous visual streams and determine when to respond (Yao et al., 2026; Wang et al., 2026). TimeActBench evaluates action selection and execution timing in real-time interactive models through their native interfaces.

2.2 Temporal and Interactive Benchmarks

TempCompass and TemporalBench evaluate temporal understanding in observed sequences, including event order, motion, duration, and frequency (Liu et al., 2024; Cai et al., 2024). StreamingBench and SVBench introduce incremental video observation and temporal dialogue, while OVO-Bench covers retrospective, real-time, and proactive responses (Lin et al., 2026b; Yang et al., 2025; Niu et al., 2025). ProactiveVideoQA and StreamReady assess answer readiness; RIVER combines retrospective memory, live perception, and proactive response; and StreamArena extends evaluation to long-horizon memory and tool use (Wang et al., 2025b; Azad et al., 2026; Shi et al., 2026; Zhang et al., 2026).

For spoken interaction, Talking Turns, Full-Duplex-Bench, and FD-Bench evaluate turn-taking, interruptions, overlap, and response latency (Arora et al., 2025; Lin et al., 2025; Peng et al., 2025). Game-Time adds explicit timing, tempo, and synchronization constraints (Chang et al., 2026). VideoFDB and OmniInteract extend evaluation to Audio-Visual agents, including conversational dynamics, response correctness, and timeliness (Mazumdar et al., 2026; Lu et al., 2026). These benchmarks establish temporal understanding and interaction timing as evaluation targets. TimeActBench brings task-specific detection signals, linguistic responses, and control operations into a common causal-streaming evaluation, using the prescribed action and its execution time as the primary units of assessment (Table 1).

3 TimeActBench

3.1 Problem Statement

TimeActBench evaluates whether and when a model executes prescribed task-specific actions under causal streaming observations. Given a task instruction q, the model observes only the stream available up to time t:

$$\mathcal{X}_{\le t} = \left\{ x^{(m)}(\tau) \mid \tau \le t,\ m \in \mathcal{M} \right\}, \qquad \mathcal{M} \subseteq \{\mathrm{A}, \mathrm{V}\}, \tag{1}$$

where A and V denote audio and video, respectively.

Motivated by prior work on online event detection and real-time interaction (Hoai & De la Torre, 2012; De Geest et al., 2016; Arora et al., 2025), we organize time-aware action into two complementary dimensions: **Perception and Detection (PD)** and **Response and Interaction (RI)** (Fig. 2).

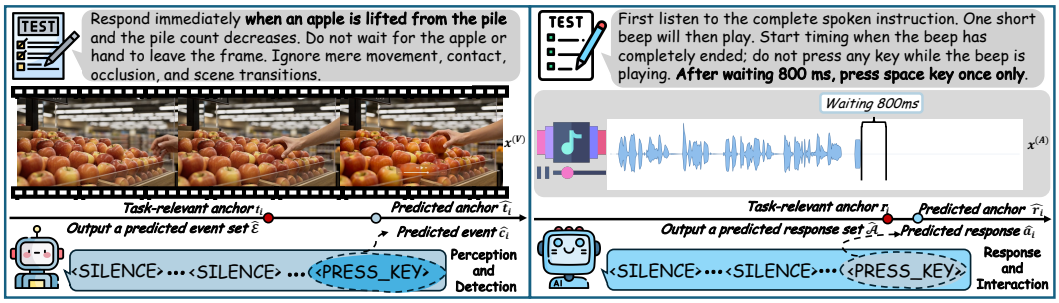


Figure 2: Illustration of the two benchmark dimensions. Left: PD detects a task-relevant video event and emits the prescribed detection action. Right: RI executes a SPACE-key action 800 ms after an auditory cue ends. Red and blue markers indicate reference and predicted action times, respectively.

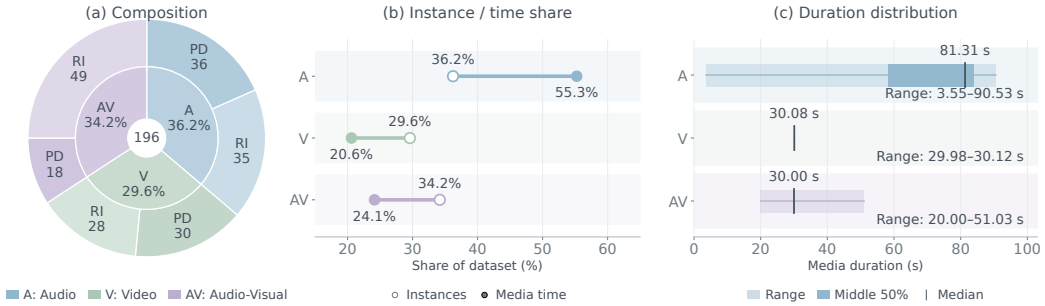


Figure 3: Dataset statistics of TimeActBench. (a) Instance distribution across modalities and task dimensions. (b) Modality shares of instances and total media duration. (c) Media-duration distributions by modality.

PD evaluates whether a model detects task-relevant events as they become observable and emits the prescribed detection action at the corresponding time. Its predicted event set is

$$\hat{\mathcal{E}} = \{(\hat{c}_i, \hat{t}_i)\}_{i=1}^{\hat{N}}, \tag{2}$$

where $\hat{N}$ is the number of predictions, $\hat{c}_i$ the event label, and $\hat{t}_i$ the detection action time. RI evaluates whether a model executes appropriate actions under temporal interaction constraints, including response timing, synchronization, and state transitions. Its predicted action set is

$$\hat{\mathcal{A}} = \{(\hat{a}_i, \hat{r}_i)\}_{i=1}^{\hat{K}}, \tag{3}$$

where $\hat{K}$ is the number of predictions, $\hat{a}_i$ the action or response, and $\hat{r}_i$ its execution time. The two dimensions indicate evaluation emphasis, not a separation of perception from control. Both dimensions score observable actions and their execution times under the prescribed triggering conditions.

3.2 Task Design and Dataset Composition

Task design is informed by research on event perception, temporal processing, and response control (Zacks, 2020; Paton & Buonomano, 2018; Friedman & Robbins, 2022). We organize tasks into seven families by their primary success conditions; definitions and family-specific motivations are provided in Appendix A. Inspired by evidence-centered design (Mislevy et al., 2003), each template specifies triggering and non-response conditions, temporal anchors, and target actions to guide construction, annotation, and scoring.

TimeActBench comprises 7 task families and 196 evaluation instances: 84 PD and 112 RI. The Audio, Video, and Audio-Visual subsets contain 71, 58, and 67 instances, respectively. Input media total 141.0 minutes, excluding instruction presentation and response-collection time. Figure 3 summarizes dataset statistics, and Appendix B.2 provides template details.

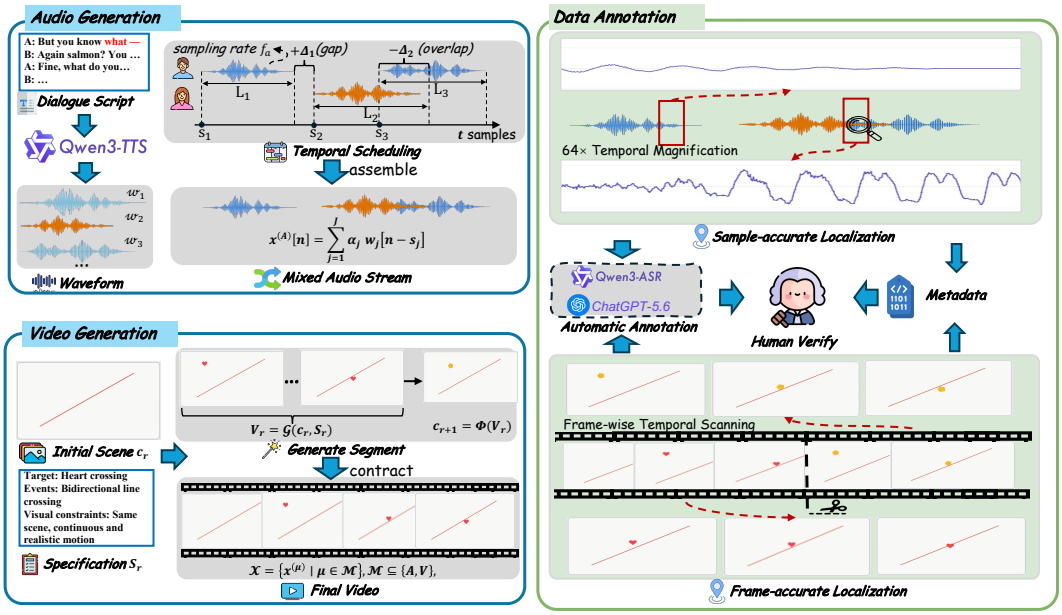


Figure 4: Data construction pipeline. Temporally controlled audio and video generation is followed by automatic annotation and human verification at sample and frame resolution.

3.3 Data Generation

To instantiate these task specifications, we generate multimodal inputs under explicit event and temporal constraints. Content generation is decoupled from temporal composition to support control over event timing. Generation metadata provide initial temporal anchors, which are checked against the final rendered media before being used as reference annotations. Fig. 4 summarizes the construction and annotation pipeline.

Audio Generation. Given a dialogue script, we independently synthesize each utterance using Qwen3-TTS (Hu et al., 2026). Let $w_j[t]$, $t = 0, \ldots, L_j - 1$, denote the waveform of the j-th utterance, where L_j is its number of audio samples. A signed interval Δ_j controls the temporal relation between consecutive utterances. Given the starting sample index s_j, the next utterance is positioned at

$$s_{j+1} = s_j + L_j + \mathrm{round}(f_a \Delta_j)\,, \tag{4}$$

where f_a is the audio sampling rate. Positive values of Δ_j produce temporal gaps, whereas negative values produce overlap. For conversational instances, the distributions of gaps and overlaps are calibrated using empirical timing statistics derived from the Fisher corpus (Cieri et al., 2004).

Given the resulting temporal schedule, the mixed audio stream is assembled at sample-level resolution as

$$x^{(A)}[n] = \sum_{j=1}^{J} \alpha_j \, w_j[n - s_j], \tag{5}$$

where J is the number of utterances and α_j is the relative gain of the j-th source; $w_j[n]$ is defined as zero outside its valid support.

Video Generation. We construct videos through event-constrained sequential generation, guided by task-specific specifications defining the initial scene, target events, temporal plans, and visual constraints. Given the initial scene condition c_1, successive video segments are generated as

$$\begin{aligned} V_r &= \mathcal{G}(c_r, \mathcal{S}_r), \\ c_{r+1} &= \Phi(V_r), \qquad r = 1, \ldots, R-1, \end{aligned} \tag{6}$$

where V_r denotes the r-th video segment, $\mathcal{G}$ is the generation procedure, c_r is the initial scene condition of segment r, and $\mathcal{S}_r$ specifies its event and scene constraints. The operator Φ extracts the terminal scene condition of V_r to condition the subsequent segment.

The validated segments are standardized and assembled into a complete video stream:

$$x^{(\mathrm{V})} = \bigoplus_{r=1}^{R} \mathcal{N}(V_r) = \{I_k\}_{k=0}^{T-1}, \tag{7}$$

where $\bigoplus$ denotes temporal concatenation, $\mathcal{N}$ denotes video standardization, and I_k is the k-th frame of the final video.

Together with the audio stream, the generated modality streams form

$$\mathcal{X} = \left\{ x^{(\mu)} \mid \mu \in \mathcal{M} \right\}, \qquad \mathcal{M} \subseteq \{\mathrm{A}, \mathrm{V}\}, \tag{8}$$

where $\mathcal{M}$ specifies the modalities present in the stimulus item. Causal access follows Eq. 1.

Data Annotation. Reference event and action annotations are defined as

$$\mathcal{E} = \{(c_\ell, \tau_\ell^*)\}_{\ell=1}^{N}, \qquad \mathcal{A} = \{(a_\ell, \rho_\ell^*)\}_{\ell=1}^{K}, \tag{9}$$

where N and K denote the numbers of reference events and actions, c_ℓ and a_ℓ their labels, and τ_ℓ^* and ρ_ℓ^* their timestamps.

Reference annotations are verified against the final rendered media rather than accepted from the generation schedule alone. Audio waveforms are magnified for sample-level localization, and videos are scanned frame by frame. Automatic annotation models independently identify task-relevant events and temporal anchors from the final media. Their outputs are compared with metadata-derived annotations, and discrepancies are resolved through human review according to the task definitions.

3.4 Evaluation Metrics

We jointly report action correctness (precision, recall, and F1) and temporal precision (MAE), following timing-sensitive interaction evaluations (Arora et al., 2025; Lin et al., 2025; 2026a; Peng et al., 2025; Défossez et al., 2024; Wang et al., 2025a; Roy et al., 2026; Cui et al., 2026).

F1 Score. F1 measures event and response correctness (Mesaros et al., 2016; Arora et al., 2025; Lin et al., 2025). Predictions are matched one-to-one with references by label, using temporal proximity only to disambiguate repeated events. Unmatched predictions and references are counted as false positives and false negatives, respectively.

Temporal precision. For correctly matched PD and RI predictions, we define the signed temporal deviations as

$$\Delta_{i\ell}^{\mathrm{PD}} = \hat{t}_i - \tau_\ell^*, \qquad \Delta_{i\ell}^{\mathrm{RI}} = \hat{r}_i - \rho_\ell^*, \tag{10}$$

where positive and negative values indicate delayed and premature predictions, respectively. We report mean absolute error (MAE) as the primary timing metric:

$$T_{\mathrm{PD}} = \frac{1}{|\mathcal{P}_{\mathrm{PD}}|} \sum_{(i,\ell)\in\mathcal{P}_{\mathrm{PD}}} \left|\Delta_{i\ell}^{\mathrm{PD}}\right|, \qquad T_{\mathrm{RI}} = \frac{1}{|\mathcal{P}_{\mathrm{RI}}|} \sum_{(i,\ell)\in\mathcal{P}_{\mathrm{RI}}} \left|\Delta_{i\ell}^{\mathrm{RI}}\right|, \tag{11}$$

where $\mathcal{P}_{\mathrm{PD}}$ and $\mathcal{P}_{\mathrm{RI}}$ denote the correctly matched prediction–reference pairs for the two dimensions. Lower values indicate more precise temporal alignment. MAE covers matched actions only and must be interpreted alongside F1; unmatched outputs and missed events affect precision and recall.

4 Experiments

4.1 Experimental Setup

Evaluated Models. The main comparison includes 15 models: 11 proprietary and four open-source systems. Proprietary models use official real-time APIs or Azure-hosted endpoints; open-source models run on dedicated NVIDIA A800 80 GB GPUs. Each model is evaluated only on its natively supported modalities, with comparisons restricted to the same subset: Audio (71), Video (58), Audio-Visual (67), or All Modalities (196). For models that support configurable reasoning effort, the main comparison uses the default reasoning effort. Sec. 4.4 examines other settings. Appendix B.1 provides model configurations.

Interaction Protocol. Task instructions precede the causally delivered media stream. Instructions are system prompts except for Moshi, which receives the same content as Qwen3-TTS speech because it lacks text input. Outputs include designated speech responses, conversational responses, speech start–stop commands, and control operations such as `press_key(key="SPACE")`; Appendix C specifies each task's action.

Network Latency. To control for transient network fluctuations, we perform an RTT probe before each evaluation run by sending WebSocket Ping frames and timing the Pong frames within the actual API session. If the measured RTT exceeds 200 ms, we retry the connection and repeat the probe until the latency falls below this threshold, and only then proceed with evaluation. Using the same procedure, we measure 20 accepted connections per endpoint and report their average RTT. Table 2 identifies cloud API baselines and reports the mean RTT; local servers follow the same measurement protocol. Action timestamps retain end-to-end latency without subtracting RTT, which measures transport rather than model inference. Appendix G.3 provides the measurements.

Human Study. We additionally conduct a human study using the same causal streaming protocol. Ten participants interact with the benchmark through a web-based evaluation interface, with tasks assigned according to the language in which each participant is proficient. Across participants, every benchmark instance is evaluated five times. Human actions are scored using the same matching and timing criteria as model actions. The study protocol and aggregate human results are provided in Appendix E.

4.2 Overall Performance and Human Comparison

Table 2 reports results by modality and over all 196 instances, with human references pooled over the same subsets.

A substantial human–model gap persists across modalities. The highest model F1 is 60.2% on Audio (GPT Live), 41.8% on Video (MOSS VL), and 33.3% on Audio-Visual (Gemini 3.8 Live), compared with human scores of 95.5%, 93.8%, and 95.2%, respectively. These correspond to gaps of 35.3, 52.0, and 61.9 percentage points. Across all 196 instances, the highest model F1 is 33.7%, compared with 95.0% for humans.

Updated models improve some dimensions without closing the gap. Gemini 3.8 Live achieves 33.3% F1 on Audio-Visual tasks, exceeding Gemini 3.1 Flash Live by 11.8 percentage points while reducing MAE from 1271 to 728 ms. Across all modalities, Gemini 3.8 Live and Gemini 2.5 Flash Live both reach 33.7% F1 after rounding, but their MAEs differ substantially: 1080 versus 1621 ms. Even the former remains well above the human MAE of 488 ms. The gains are not uniform across modalities: on Audio, Gemini 3.8 Live obtains 38.1% F1, below Gemini 3.1 Flash Live at 44.4%.

Transport latency does not explain the main timing differences. Cloud RTTs range from 18.73 to 185.31 ms. Gemini 3.8 Live and Gemini 2.5 Flash Live differ by 541 ms in overall MAE, whereas their mean RTTs differ by only 3.40 ms. Gemini 3.8 Live also has a 592 ms MAE gap to humans, compared with its 149.17 ms RTT. These scales indicate that transport alone is insufficient to explain these gaps. RTT reaches about 20% of MAE for some modality-specific results, so closely spaced timing scores should still be interpreted as end-to-end system performance.

4.3 Action Correctness and Temporal Precision

Figure 5 presents the joint F1–MAE comparison.

Low MAE does not imply reliable triggering. Qwen 3.5 Omni Plus Realtime obtains lower MAE than humans on both Video (279 versus 529 ms) and Audio-Visual (337 versus 349 ms). However, its precision is only 6.4% and 11.7%, respectively, despite recall above 83% in both modalities. This combination indicates many false-positive actions, with F1 limited to 12.0% and 20.6%. Because MAE is computed only over correctly matched actions, it does not capture the cost of unmatched outputs or missed events.

Higher precision can coexist with substantial missed actions. GPT Live achieves 78.7% precision on Audio but only 48.8% recall, missing more than half of the reference actions. On Video, MiniCPM-o 4.5 has the highest model precision (76.7%) but recalls only 22.1% of reference actions. These results reveal complementary false-activation and missed-action failure modes.

Table 2: Main results on TimeActBench. P, R, and F1 measure action correctness (%); MAE measures absolute timing error over correctly matched actions (ms). Modality-specific sections report each model on every input modality it supports. *All Modalities* reports aggregate performance over all 196 benchmark instances and therefore includes only models evaluated on Audio, Video, and Audio-Visual inputs. † marks official cloud APIs and ‡ Azure-hosted APIs; unmarked models are local. RTT is averaged over 20 connections and is not subtracted from MAE. Within each section, the best model result for each metric (P, R, F1, and MAE) is underlined; human performance is shown in bold as a reference.

Model	RTT (ms)	P ↑	R ↑	F1 ↑	MAE ↓
Audio					
GPT Live†	136.54	78.7	48.8	60.2	1954
GPT Realtime 2.1‡	184.59	37.4	71.2	49.0	1929
GPT Realtime 2.1 Mini‡	182.92	28.6	68.8	40.4	1940
GPT Realtime 2‡	185.31	25.2	74.0	37.6	1554
GPT Realtime 1.5‡	184.65	32.4	72.0	44.7	1878
Gemini 3.8 Live†	149.17	56.2	28.8	38.1	1369
Gemini 3.1 Flash Live†	151.61	72.7	32.0	44.4	1326
Gemini 2.5 Flash Live†	152.57	45.6	41.2	43.3	3248
SeedDuplex†	25.88	53.4	46.8	49.9	1808
Qwen Audio 3.0 Realtime Plus†	18.73	12.3	76.8	21.3	1089
Qwen 3.5 Omni Plus Realtime†	23.25	36.7	59.2	45.3	1560
MiniCPM-o 4.5	0.371	19.5	16.0	17.6	2478
Moshi	0.287	11.1	10.9	11.0	2229
Human	—	**96.0**	**95.1**	**95.5**	**638**
Video					
Gemini 3.8 Live†	149.17	21.2	59.7	31.3	1314
Gemini 3.1 Flash Live†	151.61	19.7	17.4	18.5	1923
Gemini 2.5 Flash Live†	152.57	19.8	53.7	28.9	1107
Qwen 3.5 Omni Plus Realtime†	23.25	6.4	87.2	12.0	279
MOSS VL	0.333	46.0	38.3	41.8	751
JoyAI VL Interaction	0.744	37.4	24.8	29.8	825
MiniCPM-o 4.5	0.371	76.7	22.1	34.4	1667
Human	—	**92.7**	**94.9**	**93.8**	**529**
Audio-Visual					
Gemini 3.8 Live†	149.17	29.3	38.6	33.3	728
Gemini 3.1 Flash Live†	151.61	29.0	17.1	21.5	1271
Gemini 2.5 Flash Live†	152.57	25.6	40.2	31.2	729
Qwen 3.5 Omni Plus Realtime†	23.25	11.7	83.2	20.6	337
MiniCPM-o 4.5	0.371	35.6	10.1	15.8	1039
Human	—	**95.4**	**95.0**	**95.2**	**349**
All Modalities					
Gemini 3.8 Live†	149.17	29.3	39.6	33.7	1080
Gemini 3.1 Flash Live†	151.61	37.4	22.4	28.0	1415
Gemini 2.5 Flash Live†	152.57	27.5	43.4	33.7	1621
Qwen 3.5 Omni Plus Realtime†	23.25	11.6	75.7	20.1	650
MiniCPM-o 4.5	0.371	30.9	14.7	19.9	1798
Human	—	**95.0**	**95.0**	**95.0**	**488**

4.4 Effect of Reasoning Effort

Figure 6 compares reasoning settings within each model: GPT Realtime models use the 71 Audio instances, while Gemini models use all 196 instances. Gemini 3.8 Live Extended Thinking is a separate variant from standard Gemini 3.8 Live in Table 2. Full results appear in Appendix D. We observe no monotonic relationship between reasoning effort and time-aware-action performance. For GPT Realtime 2.1, F1 initially increases from 44.8% at minimal to 50.5% at medium, but subsequently decreases to 44.5% at xhigh. At the same time, MAE increases from 1.81 s at minimal to 2.17 s at medium and 2.20 s at high. GPT Realtime 2 exhibits a similar correctness–timing-error pattern: increasing effort from minimal to high raises F1 from 37.5% to 41.7%, but also increases MAE from 1.41 s to 1.84 s.

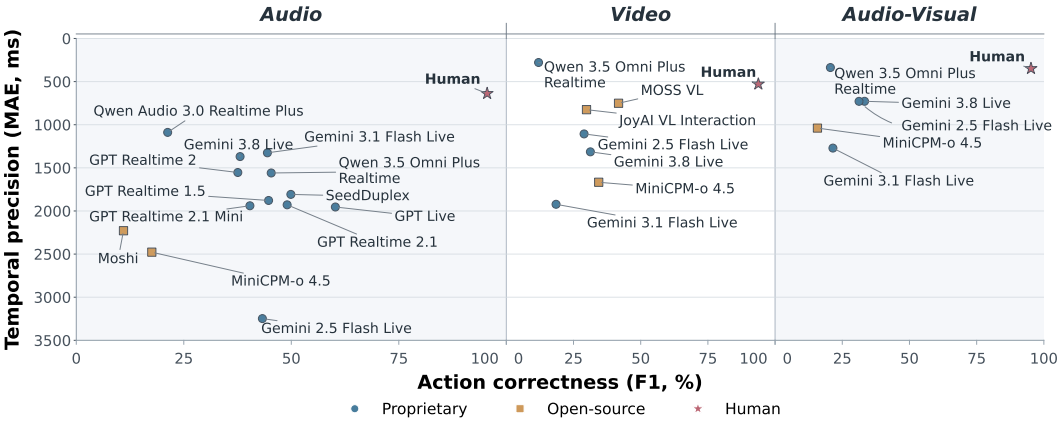


Figure 5: Joint view of action correctness (F1) and temporal precision (MAE) by input modality. The shared MAE axis is inverted; the upper right is better within each modality.

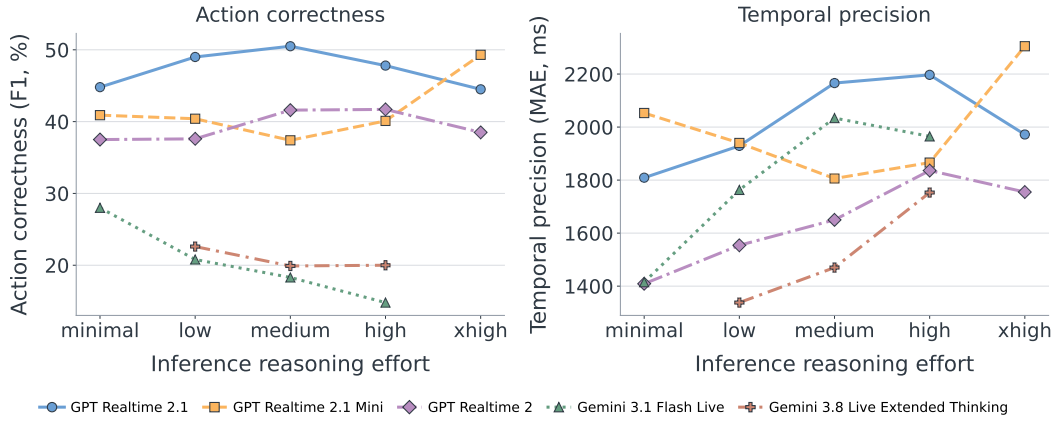


Figure 6: Effect of configurable reasoning effort on action correctness (left, F1) and temporal precision (right, MAE). Gemini 3.8 Live Extended Thinking is evaluated at low, medium, and high.

The effect is even more pronounced for Gemini 3.1 Flash Live. Its F1 decreases from 28.0% at minimal to 20.8%, 18.3%, and 14.8% at progressively higher reasoning levels, while MAE changes from 1.42 s to 1.76 s, 2.03 s, and 1.96 s, respectively. Gemini 3.8 Live Extended Thinking achieves F1 scores of 22.6%, 19.9%, and 20.0% at low, medium, and high, respectively, while MAE rises from 1.34 s to 1.75 s. GPT Realtime 2.1 Mini behaves differently again: its highest F1 is obtained at xhigh, but this setting also yields the largest timing error for the family.

Additional reasoning can improve action selection in some configurations, but the gains are not consistent across models and can come at the cost of temporal responsiveness. Reasoning settings should therefore be assessed jointly on correctness and timing.

Additional failure patterns. Beyond the false activations and missed actions above, delayed responses and output-protocol violations affect execution. Models sometimes emit explanations or non-target speech instead of the prescribed action; evaluation retains strict action extraction. Appendix F discusses these behaviors and cross-modal condition failures.

5 Conclusion

We introduced TimeActBench, a benchmark for evaluating whether and when real-time interactive models should act under causal streaming observations. Our experiments reveal a substantial overall gap between current models and humans, show that low matched-action timing error can coexist with unreliable triggering, and find no consistent benefit from increased reasoning effort. These findings highlight the need to optimize action selection and execution timing jointly. TimeActBench provides a foundation for evaluating and advancing models that translate continuous perception into reliable, precisely timed interaction. Limitations and future extensions are discussed in Appendix I.

AI USE STATEMENT

AI tools were used to assist with manuscript polishing and drafting, to support literature retrieval and discovery of related work, and to assist research ideation and execution, including the development of task designs and evaluation code. Generative models were also used to create synthetic benchmark data through the construction pipeline described in Section 3.3. Audio utterances and video segments are generated under task-specific constraints and assembled according to explicit temporal schedules. Automatic annotation models provide event and timing proposals, which are compared with generation metadata and checked against the final rendered media. Discrepancies are resolved through human review according to the task definitions. The authors reviewed and revised AI-assisted content and take responsibility for the final manuscript, methodological decisions, and reported results.

ETHICS STATEMENT

The human evaluation was conducted under the applicable institutional ethics-review procedure and approved by the Institutional Review Board (IRB). Participants were informed of the study protocol and the collection of their interaction responses before participation and provided consent according to the approved procedure. The benchmark evaluates temporal interaction capability and does not require participants to disclose sensitive personal information. Human responses are used only to construct aggregate performance references and analyze human–model performance gaps.

REPRODUCIBILITY STATEMENT

Section 3.3 describes how benchmark media are generated, temporally composed, and verified to obtain reference annotations. Section 3.4 defines the action-correctness and timing metrics. Appendix B specifies the model configurations and evaluation protocol, including supported modalities, instruction delivery, action extraction, one-to-one matching, and timestamp measurement. Appendix C lists task templates and required actions, while Appendix F.5 illustrates the scoring with task descriptions, reference times, model outputs, matching outcomes, and signed timing errors.

For timing comparisons, Appendix G reports the deployment settings and client-side stream clock used to timestamp responses. Appendix G.3 separately reports transport RTT measured by WebSocket Ping/Pong over 20 connections per endpoint; RTT is not subtracted from the reported action times. Appendix E describes participant assignment, trial coverage, and the scoring procedure for the human reference. These details document the inputs, execution conditions, and scoring rules underlying the reported results.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

REFERENCES

Alibaba. Qwen-omni-realtime technical report. Technical report, Alibaba, 2026. URL https://www.alibabacloud.com/help/en/model-studio/realtime. Accessed: 2026-09-23.

Siddhant Arora, Zhiyun Lu, Chung-Cheng Chiu, Ruoming Pang, and Shinji Watanabe. Talking turns: Benchmarking audio foundation models on turn-taking dynamics. In *International Conference on Learning Representations*, pp. 52754–52781, 2025.

Shehreen Azad, Vibhav Vineet, and Yogesh S. Rawat. Streamready: Learning what to answer and when in long streaming videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40494–40504, 2026.

Eli Brenner and Jeroen B. J. Smeets. Continuously updating one's predictions underlies successful interception. *Journal of Neurophysiology*, 2018. doi: 10.1152/jn.00517.2018.

ByteDance. Introducing seed full-duplex speech llm: Attentive listening, robust interference suppression, enabling more natural interaction. Technical report, ByteDance, 2026a. URL https://seed.bytedance.com/en/blog/introducing-seed-full-duplex-speech-llm-attentive-listening-robust-interference-suppression-enabling-more-natural-interaction. Accessed: 2026-09-23.

ByteDance. Seedrealtime: An audio-visual full-duplex llm. Technical report, ByteDance, 2026b. URL https://seed.bytedance.com/en/SeedRealtime. Accessed: 2026-09-23.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Towards fine-grained temporal understanding for multimodal video models, 2024. URL https://openreview.net/forum?id=Wto5U7q6I2.

Kai-Wei Chang, En-Pei Hu, Chun-Yi Kuan, Wenze Ren, Wei-Chih Chen, Guan-Ting Lin, Yu Tsao, Shao-Hua Sun, Hung-yi Lee, and James Glass. Game-time: Evaluating temporal dynamics in spoken language models. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16302–16306. IEEE, 2026. doi: 10.1109/ICASSP55912.2026.11464000.

Christopher Cieri, David Miller, and Kevin Walker. The fisher corpus: A resource for the next generations of speech-to-text. In *LREC*, volume 4, pp. 69–71, 2004.

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

Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *European Conference on Computer Vision*, pp. 269–284. Springer, 2016. doi: 10.1007/978-3-319-46454-1_17.

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. *arXiv preprint arXiv:2410.00037*, 2024.

Naomi P. Friedman and Trevor W. Robbins. The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*, 2022. doi: 10.1038/s41386-021-01132-0.

Minh Hoai and Fernando De la Torre. Max-margin early event detectors. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2863–2870, 2012. doi: 10.1109/CVPR.2012.6248012.

Guy Hoffman. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems*, 2019. doi: 10.1109/THMS.2019.2904558.

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

Stephen C. Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 2015. doi: 10.3389/fpsyg.2015.00731.

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. IEEE, 2025. doi: 10.1109/ASRU65441.2025.11433838.

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 *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 19447–19451, 2026a. 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. IEEE, 2026b. doi: 10.1109/ICASSP55912.2026.11463959.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024. doi: 10.18653/v1/2024.findings-acl.517.

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

Amrita Mazumdar, Seonwook Park, Rajarshi Roy, Nikhil Srihari, Shengze Wang, Yuhao Zhou, Julia Wang, Koki Nagano, and Shalini De Mello. Videofdb: Evaluating full-duplex vision-speech capabilities in conversational agents. *arXiv preprint arXiv:2605.30256*, 2026.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6):162, 2016. doi: 10.3390/app6060162.

Robert J. Mislevy, Linda S. Steinberg, and Russell G. Almond. On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 2003.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18902–18913. IEEE, 2025. doi: 10.1109/CVPR52734.2025.01761.

Uta Noppeney. Perceptual inference, learning, and attention in a multisensory world. *Annual Review of Neuroscience*, 2021. doi: 10.1146/annurev-neuro-100120-085519.

OpenAI. Introducing gpt-live. Technical report, OpenAI, 2026a. URL https://openai.com/index/introducing-gpt-live/. Accessed: 2026-09-23.

OpenAI. Gpt-realtime-2.1 technical report. Technical report, OpenAI, 2026b. URL https://developers.openai.com/api/docs/models/gpt-realtime-2.1. Accessed: 2026-09-23.

Tom Ouyang and Malini Jaganathan. Introducing gemini 3.8 live and 3.8 live extended thinking. Technical report, Google, 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-8-live-gemini-3-8-live-extended-thinking/. Accessed: 2026-09-23.

Joseph J. Paton and Dean V. Buonomano. The neural basis of timing: Distributed mechanisms for diverse functions. *Neuron*, 2018. doi: 10.1016/j.neuron.2018.03.045.

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.

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 *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. doi: 10.1109/ICASSP55912.2026.11463413.

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.

Thinking Machines Lab. Interaction models: A scalable approach to human-ai collaboration. *Thinking Machines Lab: Connectionism*, May 2026. doi: 10.64434/tml.20260511. https://thinkingmachines.ai/blog/interaction-models/.

Frederick Verbruggen et al. A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. *eLife*, 2019. doi: 10.7554/eLife.46323.

Peter Vuust, Ole A. Heggli, Karl J. Friston, and Morten L. Kringelbach. Music in the brain. *Nature Reviews Neuroscience*, 2022. doi: 10.1038/s41583-022-00578-5.

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

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63345–63354. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/wang25aw.html.

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

xAI. Introducing grok voice think fast 2.0. Technical report, xAI, 2026. URL https://x.ai/news/grok-voice-think-fast-2. Accessed: 2026-09-23.

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

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

Jeffrey M. Zacks. Event perception and memory. *Annual Review of Psychology*, 2020. doi: 10.1146/annurev-psych-010419-051101.

Xichen Zhang, Guankai Li, Yinghao Zhu, Shijian Wang, Sitong Wu, Shaozuo Yu, Meng Chu, Yuan Lu, and Jiaya Jia. Streamarena: Toward continuous, interactive, and long-horizon agentic streaming video understanding. *arXiv preprint arXiv:2608.05703*, 2026.

A Task Design and Cognitive Foundations

We organize task families by their primary success conditions and draw on cognitive research to specify their evaluation requirements. Research on event cognition motivates detecting events and boundaries in continuous activity (Zacks, 2020). The distinction between updating and response control in executive-function research motivates state updating for quantity-change monitoring and response regulation for gating and start–stop control (Friedman & Robbins, 2022). Research on temporal processing and rhythm motivates delayed responses, interval reproduction, and beat synchronization (Paton & Buonomano, 2018; Vuust et al., 2022), while multisensory perception motivates associating cues across modalities (Noppeney, 2021). Conversational timing and human–robot collaboration motivate coordinating responses with a partner's activity (Levinson & Torreira, 2015; Hoffman, 2019). Research on interception motivates coupling movement information with action timing (Brenner & Smeets, 2018); our discrete-response tasks instantiate this requirement through responses when a moving target reaches a specified location. The resulting seven families organize test instances with different response rules, as summarized in Table 3.

Table 3: Task families, behavioral requirements, and supporting literature. PD/RI indicates the evaluation emphasis; families may share perceptual and control requirements.

Task family	**Requirement and task form**	**Focus**	**References**
Event and boundary detection	Report target onsets, completions, contacts, or crossings.	PD	Zacks (2020)
Quantity-change monitoring	Monitor valid increases or decreases caused by entries and exits.	PD/RI	Friedman & Robbins (2022)
Rhythm and timing control	Respond after a delay, reproduce temporal patterns, or synchronize with beats.	PD/RI	Paton & Buonomano (2018); Vuust et al. (2022)
Cross-modal alignment and binding	Associate a spoken target with visual evidence or respond to joint conditions.	PD/RI	Noppeney (2021)
Turn-taking and interaction coordination	Detect turns, respond after a turn, or stop upon interruption.	PD/RI	Levinson & Torreira (2015); Hoffman (2019)
Response gating and start–stop control	Withhold, initiate, or stop responses according to go/no-go and control cues.	RI	Friedman & Robbins (2022); Verbruggen et al. (2019)
Motion prediction and interception	Time a discrete control response to a moving target reaching a specified location.	RI	Brenner & Smeets (2018)

Following evidence-centered design, individual tasks translate capability requirements into observable behavior (Mislevy et al., 2003). Templates specify triggering and non-response conditions, temporal anchors, and target actions, instantiated through media and instructions; annotation and scoring use consistent event definitions. Precision, recall, and F1 measure action correctness, while MAE over correctly matched actions measures temporal precision. This design supports comparisons by modality, family, and task, addressing both action outcomes and interaction timing (Hoffman, 2019).

B Detailed Evaluation Setup

B.1 Evaluated Models

We evaluate a diverse collection of proprietary and open-source real-time interactive models spanning audio, video, and Audio-Visual interaction. Proprietary systems are accessed through official real-time APIs or Azure-hosted endpoints, while open-source models are locally deployed on NVIDIA A800 GPUs with 80 GB of memory. Each open-source evaluation process is assigned a dedicated GPU to reduce interference from concurrent workloads.

Because existing real-time models expose different native input modalities, we evaluate each model only on the subset of TimeActBench supported by the interface used in our experiments. We do not convert unsupported modalities into alternative representations for the purpose of expanding model coverage. For example, an audio-only model is evaluated only on the audio subset rather than receiving automatically generated textual descriptions of video inputs. This design preserves the native real-time interaction capabilities of each system.

Table 4: Model configurations used in the main evaluation. For models with configurable reasoning effort, the listed setting corresponds to the default of the evaluated interface. The evaluation subset is determined by the modalities natively supported by the model interface used in our experiments.

Model	Access	Main Setting	Evaluation Subset
GPT Live	Official API	Default	Audio
GPT Realtime 2.1	Azure API	low	Audio
GPT Realtime 2.1 Mini	Azure API	low	Audio
GPT Realtime 2	Azure API	low	Audio
GPT Realtime 1.5	Azure API	Default	Audio
SeedDuplex	Official API	Default	Audio
Qwen Audio 3.0 Realtime Plus	Official API	Default	Audio
Moshi	Open-source	Default	Audio
MOSS VL	Open-source	Default	Video
JoyAI VL Interaction	Open-source	Default	Video
Gemini 3.8 Live	Official API	Default	All
Gemini 3.1 Flash Live	Official API	minimal	All
Gemini 2.5 Flash Live	Official API	Default	All
Qwen 3.5 Omni Plus Realtime	Official API	Default	All
MiniCPM-o 4.5	Open-source	Default	All

Table 5: Distribution of evaluation instances in TimeActBench.

	Audio	Video	Audio-Visual
PD instances	36	30	18
RI instances	35	28	49
Total instances	71	58	67

Table 4 summarizes the configurations used for the main comparison. For models exposing configurable reasoning effort, we retain the default setting of the corresponding model interface in the main results. Specifically, the default reasoning effort is `low` for GPT Realtime 2.1, GPT Realtime 2.1 Mini, and GPT Realtime 2, and `minimal` for Gemini 3.1 Flash Live. Alternative reasoning-effort settings are evaluated separately in Appendix D.

B.2 Benchmark Subsets

TimeActBench contains seven task families and 196 evaluation instances. The benchmark spans two complementary dimensions, Perception and Detection (PD) and Response and Interaction (RI), across audio, video, and Audio-Visual inputs.

Table 5 summarizes the benchmark composition.

Accordingly, an audio-only model is evaluated on 71 instances, a video-only model on 58 instances, a model supporting both video and Audio-Visual input on 125 instances, and a model supporting all three modalities on all 196 instances.

Because these subsets contain different tasks, aggregate scores obtained from different subsets are not directly comparable. The main result table reports modality-specific comparisons and an All Modalities comparison restricted to models evaluated on all 196 instances.

Figure 7 resolves the duration distribution by task template. Each PD row contains six instances and each RI row seven. The shared horizontal scale makes the near-constant video durations and the broader variation among audio tasks directly comparable.

B.3 Interaction Protocol

Each evaluation instance consists of an instruction phase followed by a scored streaming-interaction phase.

During the instruction phase, the model receives a task description specifying the temporal condition to monitor and the action that should be produced when that condition is satisfied. The scored media stream is then delivered causally in real time. At any point during evaluation, the model has access

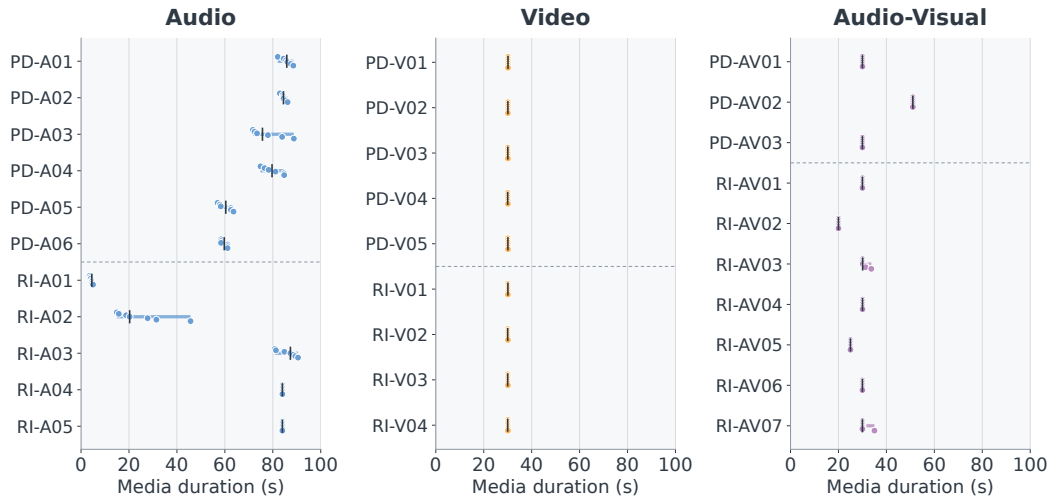


Figure 7: **Media durations across individual task templates.** Each point is one evaluation instance, with a small vertical offset to separate coincident observations. Colored segments span the minimum and maximum durations within each task template; dark ticks indicate medians. Dashed horizontal lines separate PD and RI tasks. Task identifiers follow the dataset release. All panels use the same duration scale.

only to observations that have already occurred in the stream; future observations are never exposed in advance.

For interfaces supporting textual instructions, the task description is provided through the corresponding instruction or system-prompt channel before media playback begins. The instruction phase itself is excluded from temporal scoring.

Moshi does not expose a textual instruction channel in the interaction configuration used in our experiments. For Moshi, we therefore synthesize the same task instruction using Qwen3-TTS and present the resulting speech before the scored media begins. The synthesized instruction is not included in the scored stream.

Once evaluation begins, models are expected to continuously process incoming observations and either remain inactive or emit the prescribed action whenever the task-specific temporal condition is satisfied.

B.4 Action Interface

We design task outputs to be compact, externally observable, and straightforward to timestamp. Depending on the task, the required output may take one of several forms:

- a designated spoken token, such as “hit”;
- a natural conversational response;
- stopping or resuming an ongoing spoken response;
- a discrete control action, such as `press_key(key="SPACE")`; or
- one of several task-specific discrete control choices.

Although some actions are represented through textual model outputs before being mapped to executable operations, we evaluate the resulting task-specific action rather than the linguistic quality of the generated text. The benchmark therefore focuses on whether the prescribed operation is produced and when it occurs.

B.5 Action Extraction and Matching

Model outputs are converted into a sequence of timestamped actions before scoring. For tasks requiring a designated spoken token, the evaluator detects occurrences of the target response in the model output stream. For control tasks, the corresponding function or action invocation is recorded directly. Outputs that do not correspond to a valid task action are not considered correct predictions.

Predicted actions are matched to reference actions using the task-specific temporal matching procedure defined by the benchmark. Each prediction can match at most one reference action, and each reference action can match at most one prediction.

A matched prediction is counted as a true positive (TP). An unmatched model action is counted as a false positive (FP), while an unmatched reference action is counted as a false negative (FN).

We compute precision, recall, and F1 as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

and

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

These metrics jointly characterize whether a model reliably determines when an action is warranted. In particular, precision penalizes unnecessary actions, while recall penalizes missed required actions.

B.6 Timing Measurement

For every correctly matched prediction-reference pair, we measure the absolute temporal error between the predicted action time and the reference action time.

Let t_i^{pred} denote the timestamp of a correctly matched model action and t_i^{ref} the corresponding reference timestamp. The absolute temporal error is

$$e_i = \left| t_i^{\text{pred}} - t_i^{\text{ref}} \right|.$$

Temporal precision is measured using mean absolute timing error (MAE):

$$\text{MAE} = \frac{1}{N_{\text{match}}} \sum_{i=1}^{N_{\text{match}}} e_i,$$

where N_{match} is the number of correctly matched actions.

Lower MAE indicates that correctly produced actions occur closer to their intended temporal locations.

Importantly, MAE is defined only over correctly matched actions. It therefore cannot determine whether a model acts too frequently or fails to act when required. For example, a model that repeatedly emits the target action may obtain a low MAE whenever one of its many outputs happens to fall close to the reference timestamp, despite producing numerous false positives. For this reason, MAE must be interpreted jointly with precision, recall, and F1.

We retain all valid benchmark instances when computing MAE. In particular, tasks involving long delays or extended waiting periods are intentional components of TimeActBench rather than timing outliers and are not removed from evaluation.

C Task Templates

The seven task families are instantiated through task templates that specify concrete response rules. Each PD task template contains six instances, while each RI task template contains seven instances, resulting in 196 evaluation instances in total.

Table 6 summarizes the temporal behavior and required action associated with each task template.

Table 6: Task taxonomy of TimeActBench. "Hit" denotes a designated spoken response, while SPACE and numeric keys denote discrete control actions.

ID	Modality	Temporal Behavior	Required Action	#
PD-A01	A	Speaker-turn onset	Say "hit"	6
PD-A02	A	Interruption onset	Say "hit"	6
PD-A03	A	Target-word completion	Say "hit"	6
PD-A04	A	Quantifier detection	Say "hit"	6
PD-A05	A	Target sound onset	Say "hit"	6
PD-A06	A	Sound-pattern completion	Say "hit"	6
PD-V01	V	Region entry/exit transition	SPACE	6
PD-V02	V	Goal-line crossing	SPACE	6
PD-V03	V	Seating completion	SPACE	6
PD-V04	V	Vehicle line crossing	SPACE	6
PD-V05	V	Ball–racket contact	SPACE	6
PD-AV01	AV	Visible-speaker turn transition	SPACE	6
PD-AV02	AV	Audio-Visual category match	SPACE	6
PD-AV03	AV	Audio-Visual semantic match	SPACE	6
RI-A01	A	Rhythm reproduction	Say "hit" rhythmically	7
RI-A02	A	Delayed response	Say "hit" after a prescribed delay	7
RI-A03	A	Conversational turn response	Spoken response	7
RI-A04	A	Interruption handling	Stop/resume speech	7
RI-A05	A	Spoken start–stop control	Stop/resume speech	7
RI-V01	V	Moving-object interception	SPACE	7
RI-V02	V	Release-state response	SPACE	7
RI-V03	V	Go/no-go visual response	SPACE	7
RI-V04	V	Grasp-completion response	SPACE	7
RI-AV01	AV	Beat-aligned target response	SPACE	7
RI-AV02	AV	Multimodal rhythm reproduction	Say "hit" rhythmically	7
RI-AV03	AV	Speech-conditioned visual interception	SPACE	7
RI-AV04	AV	Orientation-completion response	SPACE	7
RI-AV05	AV	Cue-conditioned rhythmic control	Key 1/2/3	7
RI-AV06	AV	Cross-modal start–stop control	SPACE	7
RI-AV07	AV	Audio-Visual conjunction response	SPACE	7

D Reasoning Effort

Several proprietary real-time models expose configurable reasoning effort. We compare the reasoning-effort settings listed in Table 7 for GPT Realtime 2.1, GPT Realtime 2.1 Mini, GPT Realtime 2, Gemini 3.1 Flash Live, and Gemini 3.8 Live Extended Thinking.

Table 7 reports the complete results. GPT Realtime models use the 71 Audio instances; Gemini models use all 196 instances. Reasoning-effort comparisons are made within each model.

Across model families, increasing reasoning effort does not produce monotonic improvements in time-aware-action performance. Some configurations improve action correctness at intermediate reasoning levels while simultaneously increasing timing error. Other models deteriorate in both dimensions as reasoning effort increases.

These results suggest that additional inference computation does not automatically translate into stronger temporal intelligence. In real-time interaction, reasoning must not only improve the decision itself but also preserve the temporal responsiveness required to execute that decision at the appropriate moment.

E Human Evaluation

E.1 Human Study Protocol

We conduct a human evaluation under the same causal streaming setting used for model evaluation. Ten participants complete benchmark instances through a web-based real-time interaction interface.

Participants receive the task instruction before the scored stream begins and subsequently observe or listen to the same temporally paced media presented to models. They respond using the action interface defined for the corresponding task.

Tasks are assigned according to the language in which each participant is proficient. Across participants, every benchmark instance receives five human trials.

Table 7: Reasoning-effort results (F1 in %, MAE in ms).

Model	Reasoning Effort	F1 ↑	MAE ↓
GPT Realtime 2.1	minimal	44.8	1809
	low	49.0	1929
	medium	50.5	2166
	high	47.8	2197
	xhigh	44.5	1972
GPT Realtime 2.1 Mini	minimal	40.9	2053
	low	40.4	1940
	medium	37.4	1806
	high	40.1	1866
	xhigh	49.3	2305
GPT Realtime 2	minimal	37.5	1409
	low	37.6	1554
	medium	41.6	1650
	high	41.7	1836
	xhigh	38.5	1755
Gemini 3.1 Flash Live	minimal	28.0	1415
	low	20.8	1763
	medium	18.3	2034
	high	14.8	1965
Gemini 3.8 Live Extended Thinking	low	22.6	1338
	medium	19.9	1470
	high	20.0	1753

Table 8: Human performance on TimeActBench by modality. MAE is reported in milliseconds.

Modality	P ↑	R ↑	F1 ↑	MAE ↓
Audio	95.96	95.12	95.54	638
Video	92.66	94.90	93.77	529
Audio-Visual	95.42	95.00	95.21	349

Human responses are scored using the same action-matching and temporal-evaluation procedure used for model outputs.

E.2 Human Performance

Table 8 reports aggregate human performance for each modality.

The modality-specific human references in the main comparison are obtained directly from the corresponding trials in Table 8. For the *All Modalities* result, precision, recall, and F1 are recomputed by pooling TP, FP, and FN across Audio, Video, and Audio-Visual trials.

Because MAE is defined only over correctly matched actions, the pooled MAE is computed by weighting each modality-specific MAE by its corresponding number of matched actions.

This yields the human references reported in the main paper: 95.5% F1 and 638 ms MAE for Audio, 93.8% and 529 ms for Video, 95.2% and 349 ms for Audio-Visual, and 95.0% and approximately 488 ms over all 196 instances.

F Qualitative Failure Analysis

Aggregate metrics characterize overall performance but do not fully expose the interactive behaviors that produce those scores. We therefore visualize representative model traces on a shared temporal axis.

Each timeline displays reference actions together with model outputs and indicates correctly matched actions, false positives, missed actions, and temporally displaced responses.

We focus on four recurrent failure patterns.

F.1 Over-Triggering

Some models repeatedly produce the target action even when the required temporal condition is absent.

This behavior produces many false positives and therefore low precision. At the same time, frequent action generation increases the probability that at least one prediction will occur close to a reference event. Consequently, an over-triggering model may obtain relatively low MAE on its matched actions despite failing to reliably determine whether an action is warranted.

This failure mode illustrates why temporal precision should never be interpreted independently of action correctness.

F.2 Missed or Delayed Actions

A second class of failures occurs when a model fails to act at a required event or responds only after substantial additional evidence has accumulated.

Such behavior reduces recall and can also increase timing error for successfully matched responses.

These failures are particularly important for tasks in which action utility rapidly decreases after the target event, including interruption handling, synchronization, moving-object interception, and state-transition responses.

F.3 Cross-Modal Condition Failures

Audiovisual tasks frequently require the model to act only when evidence from both modalities satisfies a prescribed temporal condition.

A model may correctly identify an acoustic event or visual event in isolation while failing to enforce the required conjunction between them. Such failures reveal that successful unimodal event detection does not necessarily imply reliable cross-modal temporal binding.

F.4 Output-Protocol Violations

We additionally observe cases in which models generate explanations, timestamps, or other linguistic content despite being instructed to emit only the prescribed task action.

These cases do not necessarily correspond to failures of temporal perception itself. However, they are important in real interactive environments because an action must not only be internally recognized but also translated into the required externally executable behavior.

We therefore retain strict action-extraction rules and report representative protocol violations qualitatively rather than introducing a separate language-quality metric.

F.5 Example Items and Scored Responses

The examples below show an Audio, Video, and Audio-Visual item from the sampled evaluation records, with task instructions summarized in English. The response column reproduces the exported *scored value*, i.e., the action extracted for evaluation rather than a full conversational transcript. Times are measured from media onset. Match status is retained from the scoring export; "hit" indicates a matched action and "false alarm" an unmatched output. Signed latency is response time minus target time, with positive values indicating late responses. An unmatched output has no assigned target or latency. Values are rounded only for display.

Delayed audio response (RI-A02/T1). After the short cue ends, wait 500 ms before producing the designated response `hit`. The target response is `hit` at 10.752917 s.

Table 9: Scored responses for RI-A02/T1. Response and target times are in seconds; signed latency is in milliseconds. All recorded scored responses are shown for the listed models.

Model	Response	Time	Match status	Target	Latency
GPT Live	`hit`	11.5935	hit	10.7529	+840.6
Gemini 3.8 Live	`hit`	11.0459	hit	10.7529	+293.0
MiniCPM-o 4.5	`hit`	10.3463	false alarm	—	—

Gemini 3.8 Live and GPT Live produce matched responses with different delays. The MiniCPM-o output is marked as a false alarm in the scoring record and is therefore not assigned a timing error.

Visual state transition (PD-V03/T1). Press SPACE when the patient first sits on the wheelchair seat in each scene. Do not respond during standing, turning, support, or descent. The target actions are `SPACE` at 10.250000 and 27.875000 s, corresponding to the first seat contact that persists in subsequent frames.

Table 10: Scored responses for PD-V03/T1. Response and target times are in seconds; signed latency is in milliseconds. Qwen rows are excerpts; all recorded scored responses are shown for the other listed models.

Model	Response	Time	Match status	Target	Latency
Gemini 3.8 Live	SPACE	31.4507	hit	27.8750	+3575.7
	SPACE	33.5621	false alarm	—	—
MiniCPM-o 4.5	SPACE	28.5051	hit	27.8750	+630.1
MOSS VL	SPACE	10.2498	false alarm	—	—
	SPACE	11.2160	hit	10.2500	+966.0
	SPACE	27.2309	false alarm	—	—
	SPACE	28.2035	hit	27.8750	+328.5
Qwen 3.5 Omni Plus Realtime	SPACE	1.5851	false alarm	—	—
	SPACE	2.6929	false alarm	—	—
	SPACE	10.5886	hit	10.2500	+338.6
	SPACE	27.8876	hit	27.8750	+12.6

Qwen 3.5 Omni Plus Realtime produces 40 scored outputs for this item, of which two match the target events and 38 are false alarms. Its small timing error on the matched actions therefore coexists with substantial over-triggering.

Audio-Visual conditional response (RI-AV07/T2). Press SPACE only when a left arrow and a double-click sound jointly establish the target combination. Withhold responses when either cue is absent or the combination is incorrect. The target actions are `SPACE` at 5.166667 and 27.583333 s.

Table 11: Scored responses for RI-AV07/T2. Response and target times are in seconds; signed latency is in milliseconds. All recorded scored responses are shown for the listed models.

Model	Response	Time	Match status	Target	Latency
Gemini 2.5 Flash Live	SPACE	6.7555	hit	5.1667	+1588.8
	SPACE	28.7811	hit	27.5833	+1197.7
Gemini 3.8 Live	SPACE	1.8483	false alarm	—	—
	SPACE	3.9067	false alarm	—	—
Gemini 3.1 Flash Live (minimal)	SPACE	17.9859	false alarm	—	—
	SPACE	20.1664	false alarm	—	—
	SPACE	30.2494	hit	27.5833	+2666.0

Gemini 2.5 Flash Live matches both target events. Gemini 3.8 Live produces two false alarms and no matched action, while Gemini 3.1 Flash Live matches only the second target after two false alarms.

G Runtime and Reproducibility

G.1 Open-Source Model Deployment

Open-source models are evaluated on NVIDIA A800 GPUs with 80 GB of memory. Each evaluation process is assigned a dedicated GPU. Model-specific inference servers and streaming interfaces are initialized before benchmark playback begins so that startup cost is excluded from action timing.

All benchmark media are replayed according to their original temporal schedule rather than being provided as offline batches. This ensures that locally deployed models receive observations causally as the interaction unfolds.

Table 12: WebSocket Ping/Pong RTT over 20 connections per endpoint. All values are in milliseconds. Extended Thinking denotes the separately measured Gemini 3.8 endpoint configuration.

Model / endpoint	Access	Mean	Median	P95
GPT Live	Official API	136.54	134.06	182.87
GPT Realtime 2.1	Azure API	184.59	186.81	195.08
GPT Realtime 2.1 Mini	Azure API	182.92	180.92	195.28
GPT Realtime 2	Azure API	185.31	184.94	195.16
GPT Realtime 1.5	Azure API	184.65	185.59	196.97
Gemini 3.8 Live	Official API	149.17	146.19	160.49
Gemini 3.1 Flash Live	Official API	151.61	152.95	161.70
Gemini 2.5 Flash Live	Official API	152.57	152.02	185.45
SeedDuplex	Official API	25.88	20.38	73.58
Qwen Audio 3.0 Realtime Plus	Official API	18.73	14.35	67.90
Qwen 3.5 Omni Plus Realtime	Official API	23.25	16.29	69.11
Gemini 3.8 Live Extended Thinking	Official API	153.27	153.22	182.97
MiniCPM-o 4.5	Local	0.371	0.357	0.487
MOSS VL	Local	0.333	0.342	0.386
Moshi	Local	0.287	0.284	0.327
JoyAI VL Interaction	Local	0.744	0.751	0.857

G.2 Proprietary Real-Time APIs

GPT Realtime variants are evaluated through Azure-hosted endpoints; other proprietary models use their providers' official real-time interfaces. The same client-side evaluation framework is used to stream observations, record outputs, and timestamp model actions.

End-to-end response timing from proprietary APIs includes service-side scheduling, transport latency, and other components of the deployed interaction stack. We intentionally retain these components because TimeActBench evaluates realized interactive behavior rather than isolated internal model computation.

Accordingly, timing results should be interpreted as end-to-end performance under the reported evaluation environment.

G.3 WebSocket Transport Latency

For each endpoint, we establish 20 real WebSocket sessions and measure elapsed time from sending a Ping frame to receiving its corresponding Pong frame. Table 12 reports the mean, median, and 95th-percentile RTT over 20 accepted connections for each endpoint. Measurements characterize the transport path used in evaluation, including the endpoint's handling of Ping/Pong frames. They are not inference latencies or estimates of one-way delay. We report the measured end-to-end action times without subtracting RTT or half of RTT.

G.4 Timing Consistency

All timestamps used for scoring are recorded relative to the benchmark stream clock maintained by the evaluation client. The same timestamping implementation is used across models within each action type.

H Additional Implementation Details

H.1 Instruction Consistency

Task instructions describe the same underlying condition and required action across evaluated systems. Interface-specific formatting is used only when required by a model API.

We avoid introducing model-specific hints about the location or timing of target events. Instructions specify the response rule without revealing the future occurrence times of target events in the scored stream.

H.2 Unsupported Modalities

If a model interface does not natively support a benchmark modality, the corresponding task is excluded from that model's evaluation.

We do not use modality conversion as a substitute for native perception. For example, video is not converted into textual captions for an audio/text model, and audio is not transcribed before being provided to a visual model.

This choice avoids conflating the temporal capability of the evaluated model with that of an external preprocessing system.

H.3 Invalid and Extra Outputs

Outputs that cannot be mapped to the required task-specific action are not counted as correct predictions.

If a model emits multiple valid actions where only one is required, the benchmark matching procedure can match at most one prediction to the corresponding reference event. Remaining unmatched actions are treated as false positives.

This scoring rule discourages models from improving recall through indiscriminate repeated triggering.

I Limitations

TimeActBench uses fixed media streams and prescribed actions to enable reproducible timing comparisons. Model actions do not alter subsequent observations, so generalization to closed-loop interactions with adaptive partners or physical environments remains to be established. Reported timing reflects the full deployed system, including inference, service scheduling, and transport; Ping/Pong measurements characterize transport but do not isolate these components. Rankings between systems with similar timing scores may therefore vary with deployment conditions. The human reference comprises ten participants, with five trials per instance, and provides a comparison under the benchmark protocol; larger and more diverse samples are needed to characterize variation across users. Future extensions can address these boundaries through action-dependent environments, repeated evaluations across deployments, and broader participant recruitment.