OMNIASSISTBENCH: ASSISTANT-STYLE INTERACTION BENCHMARK FOR OMNI-LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent omni-modal large language models (Omni-LLMs) show great potential as real-time video assistants,which continuously perceive environments and guide users to achieve specific goals. Unlike traditional passive video understanding, interactive assistants should actively combine visual states, user goals, and prior knowledge to provide effective help. Evaluating this is rather challenging, as the model's unpredictable response dynamically changes the user's subsequent actions, which static offline datasets cannot accommodate. To address this bottleneck, we introduce **OmniAssistBench**. To solve the issue of diverging interaction paths (where the same user goal can be achieved through various methods), we provide models with predefined priors derived from the source video, requiring them to guide users along the exact same routes. Since real interaction videos are rare, we construct the dataset by reverse-engineering existing Internet videos. We deduce logical user goals and segment the videos into multi-turn clips to simulate continuous interactions. This rigorous pipeline required **over 1000 expert person-hours** to build the dataset. Results show that the proprietary Gemini-3-Pro reaches 66.4 out of the max point of 100, while the open-source Qwen3-Omni-Instruct achieves 51.2. Although current models generally understand user inputs, they frequently provide incorrect or incomplete answers. Specifically, they struggle with visual prompts (e.g., hand gestures), fail to maintain historical context during multi-turn interactions, and fail to delay response until the target event. Results indicate room for improvement before models can become reliable assistants.

1 INTRODUCTION

Emerging omni-modal large language models (Omni-LLMs) (Li et al., 2025b; Fu et al., 2025b; Xu et al., 2025a;b; Google DeepMind, 2025) are evolving from passive multimodal predictors into real-time interactive video assistants. Unlike traditional video understanding, which passively analyzes given video content, an interactive assistant must actively combine the current visual state, user goals, and prior knowledge to continuously guide users toward specific objectives. However, assessing these models under realistic, assistant-style video chat scenarios remains challenging.

The primary difficulty in evaluating interactive understanding lies in dataset construction. As shown in Fig. 1, in traditional video understanding benchmarks such as Lin et al. (2026); Mangalam et al. (2023); Li et al. (2024a); Fu et al. (2025a), content of the test videos in the following turns do not need to change according to the model's reply in previous turns. This allow researchers to gather large amount of Internet videos and add static question-answer pairs. In contrast, in real-world multi-turn interaction scenarios, the assistant's response influences the user's subsequent actions, resulting in diverging interaction paths: users might achieve the same goal through different methods depending on the assistant's guidance. Static offline datasets cannot dynamically adjust video inputs to match unpredictable model responses, nor can researchers exhaust all possible interaction paths to prepare corresponding videos. This makes standard evaluation pipelines invalid for interactive scenarios.

To tackle this evaluation bottleneck, we introduce OmniAssistBench. To solve the issue of diverging interaction paths, we propose a data construction methodology based on prior knowledge summarized from the video content. Specifically, we provide models with explicit prior knowledge (e.g., a standard procedural path) and require them to guide users strictly along this unique route. This prevents user actions from diverging and reduces the evaluation's reliance on the model's internal

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.

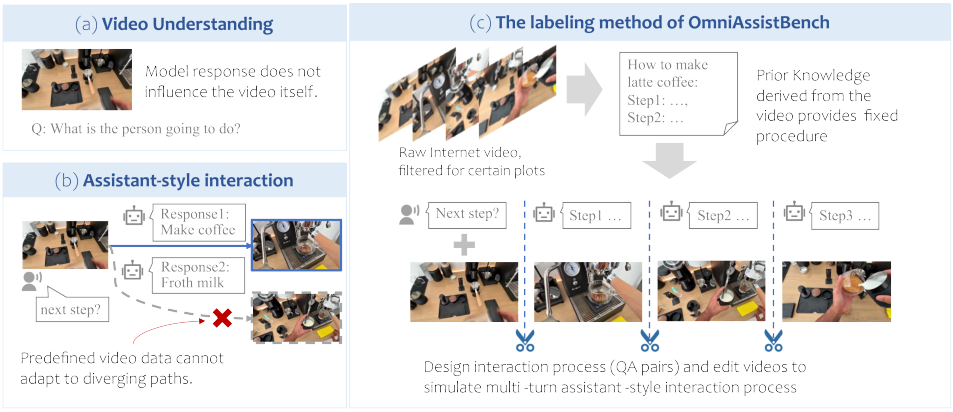


Figure 1: (a) In traditional video understanding evaluation, model's responses do not change the content of the test video. (b) Assistant models may guide the user to achieve one goal through different paths (e.g., Both Response 1 and Response 2 shown in the figure are valid, but test data only contains the consequent video of Response 1.) (c) To address this problem of path diversity, we manually filters for videos with certain plots and deduces prior knowledge from the video content to enforce a fixed interaction path. Based on the paths, experts then design detailed interaction turns and edit the videos accordingly to build human-assistant interaction recordings.

knowledge. Furthermore, since direct recordings of complete human-model interactions are rare, we construct the dataset by reverse-engineering existing Internet videos. We first filter suitable source videos, logically deduce the user goals and path priors from the video content, and then segment the videos into multi-turn clips to simulate the step-by-step interaction between a user and an assistant.

Aiming at building a comprehensive benchmark based on this methodology, OmniAssistBench features a two-tier evaluation framework: the Basic tier targets foundational perception skills which are critical for interaction, while the Advanced tier evaluates complex, goal-oriented tasks using plot templates summarized from everyday scenarios. To better reflect model performance in real world settings, we also film 3 cases with an average of 15 interaction turns. Each case has its plot specifically designed to target the combinations of a group of the abilities evaluated in the benchmark. To mimic natural usage, user instructions are directly embedded into the test videos as audio, paired with an open-ended scoring metric to prevent multiple-choice shortcuts. Due to the complex script design and heavy manual video editing, building this benchmark took over 1000 expert person-hours.

Using our rigorous scoring protocol, which penalizes redundant and hallucinated information, we evaluate state-of-the-art Omni-LLMs. The proprietary Gemini-3-Pro (Google DeepMind, 2025) reaches 66.4 out of 100 points, while the open-source Qwen3-Omni-Instruct (Xu et al., 2025b) achieves 51.2. Our evaluation highlights three major bottlenecks in current models: (1) Deficiency in following visual prompts (e.g., hand gestures); (2) Limitations in long-term memory, where context limits are quickly exhausted within minutes; and (3) Failure in delayed response, as models often reply immediately with unrelated captions instead of waiting for the requested visual plot.

In summary, our main contributions are as follows:

- We propose a novel dataset construction pipeline specialized for assistant-style interaction scenarios. Based on existing Internet videos, We derive priors to provide fixed interaction paths and simulate assistant-style interaction processes by reverse-engineering.
- We introduce OmniAssistBench which features a two-tier framework To better reflect model performance in complicated real-world scenarios which usually require multiple abilities, we also designed and filmed 3 specific cases from scratch.
- We conduct extensive evaluations on state-of-the-art Omni-LLMs, uncovering shortcomings in multimodal instruction following, long-term memory, delayed response, and cross-turn context management. These findings provide potential directions for the future development of real-world assistant models.

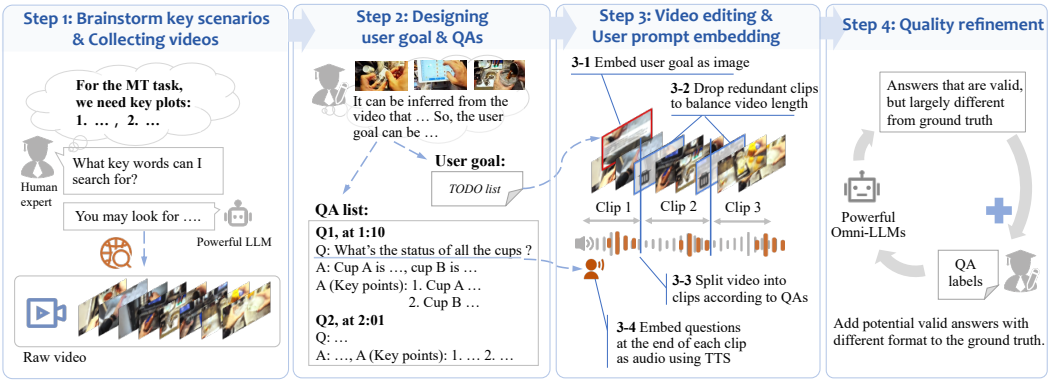


Figure 2: The 4-step data construction process of OmniAssistBench, taking the Multitask Tracking (MT) task as an example. The requirements of designing logically valid interaction process and heavy video editing make it difficult to replace human labor with MLLM-based automatic pipeline.

2 Data Construction

To reflect real-world assistant capabilities, we implement a multi-stage data construction pipeline relying on human experts. We opt for manual annotation over automated pipelines because: (1) our tasks target subtle, non-salient video details often missed by LLM-based captions; and (2) the Advanced-tier tasks center on user goals, requiring human judgment to ensure logical alignment with the video content.

2.1 Data Format

A typical Question-Answer (QA) pair in OmniAssisBench consists of a video with a question embedded in it, a ground truth answer paragraph, and a few key point phrases of the ground truth answer. We adopt open-ended questions and a multi-turn format to mimic real time online interactions.

Open-Ended Questioning: We adopt the open-ended question format over multiple-choice questions to mitigate the risk of information leakage. We observe that when solving multiple-choice questions, models tend to focus more on the options: Models usually first look up the video for cues of the objects mentioned in the options, then predict whether they relate to the questions or not. However, this way of thinking contradicts the real application cases where no options are available: Models need to first comprehend the questions, then actively look for relative cues in the video. Consequently, the predefined options in multiple-choice questions may provide shortcuts by listing objects that are potentially relevant.

Offline Simulation of Online Interaction: To simulate online temporal causality, we embed user prompts at the very end of the input video as audio, or as picture-in-picture for Non-audio Prompt Following tasks. This mimics real-time inference, where forward passes initiate upon prompt detection. For multi-turn interactions, the start of the current video segment aligns with the timestamp of the previous turn's conclusion to ensure temporal continuity.

2.2 Data Construction Pipeline

Data construction process consists of four phases: scenario design & video collecting, QA design, video editing & user prompt embedding, and quality refinement. The pipeline is illustrated in Fig. 2.

Scenario design & video collecting: For each task category, human experts collaborate with LLMs to brainstorm key scenarios that challenge specific model capabilities. We then source raw footage mainly from YouTube[1], supplemented by curated clips from existing datasets without directly using

[1] https://www.youtube.com

Table 1: Task list and statistics of the dataset grouped by second-level task categories, including video counts, QA pairs, turns of interaction, and video duration metrics. Note that the number of turns denotes the annotation target rather than exact averages, as actual turn numbers depends on the video.

Tier	Major Tasks	Sub Tasks (Abbr.)	# Vid.	# QA	#Turns	Avg. Vid. len (s)
Basic Tier	Social Perception	Identity Identification (II), Addressee Identification (AI), Complex Emotion Understanding (CE)	67	67	1	49.46
	Temporal Perception	Event Retrieval (ER), Appearance Order Perception (AO), Dynamic Counting (DC)	47	47	1	77.03
	Referential Perception	Action Reference Perception (AR), Linguistic Reference Perception (LR)	49	49	1	40.86
	Non-audio Prompt Following	Gesture-based Prompt Following (GPF), OCR-based Prompt Following (OPF)	30	76	2	149.84
Advanced Tier	Context-aware Response	Context-aware Response (CR)	20	20	1	119.62
	Proactive Response	Single Event-triggered Response (SER), Multi Event-triggered Response (MER)	32	133	4	209.03
	Process Tracking	Step Tracking (ST), Checklist Tracking (CT), Multitask Tracking (MT)	52	246	5	249.21
Real World Cases			3	47	15	1016.13
ALL			300	685	-	182.25

Table 2: Distribution of video topics across task tiers. Here, *# Video* denotes the number of unique videos, and *# QA* represents the total number of question-answer pairs. The number of QA pairs exceeds the video count since many videos feature multi-turn interactions.

Tier	Metric	Lifestyle & Home	Food & Cooking	Travel & Outdoors	Tech & DIY	Fitness & Health	Entertainment & Gaming	Education & Info	**Total**
Basic	# Video	34	17	8	11	10	73	40	193
	# QA	41	37	10	11	10	91	41	241
Advanced	# Video	12	22	8	24	4	14	23	107
	# QA	29	96	30	110	18	55	108	446
Total	# Video	46	39	16	35	14	87	63	**300**
	# QA	70	133	40	121	28	146	149	**687**

the original caption (Sener et al., 2022; Grauman et al., 2022; Pan et al., 2025; Rossetto et al., 2025; Tang et al., 2019; Poria et al., 2019; Ray et al., 2022; Oh et al., 2011).

Since we search for videos based on detailed plots (e.g., an social scene with more than 3 people constantly appearing in the frames, and one of them is talking to another) rather than broad topics or keywords, the video collection phase can be time-consuming. We also include several home-filmed samples for scenarios where the required plots are difficult to locate online. While certain plots naturally appear more frequently in specific video categories (e.g., multi-person communication in interviews), we try our best to ensure the collected videos uniformly cover common daily-life scenarios. Table 2 displays the video topics of benchmark.

QA design: Experts design QA pairs based on both the source video and the task type. For Basic-tier tasks, questions ask about the objects or events in the video (e.g., "What did the person say when they turned on the coffee machine?"). In contrast, for Advanced-tier tasks, questions are based on user goals and prior knowledge summarized from the video. User goals should be partly independent from the source videos, and the target objects are usually not direly mentioned in the questions (e.g., the question "What is the next step for making this cup of coffee?" makes sense for most videos about making coffee, and does not mention any detailed coffee machines). Prior knowledge ensure fix interaction path of how models are expected to assist user to achieve their goals. To satisfy this

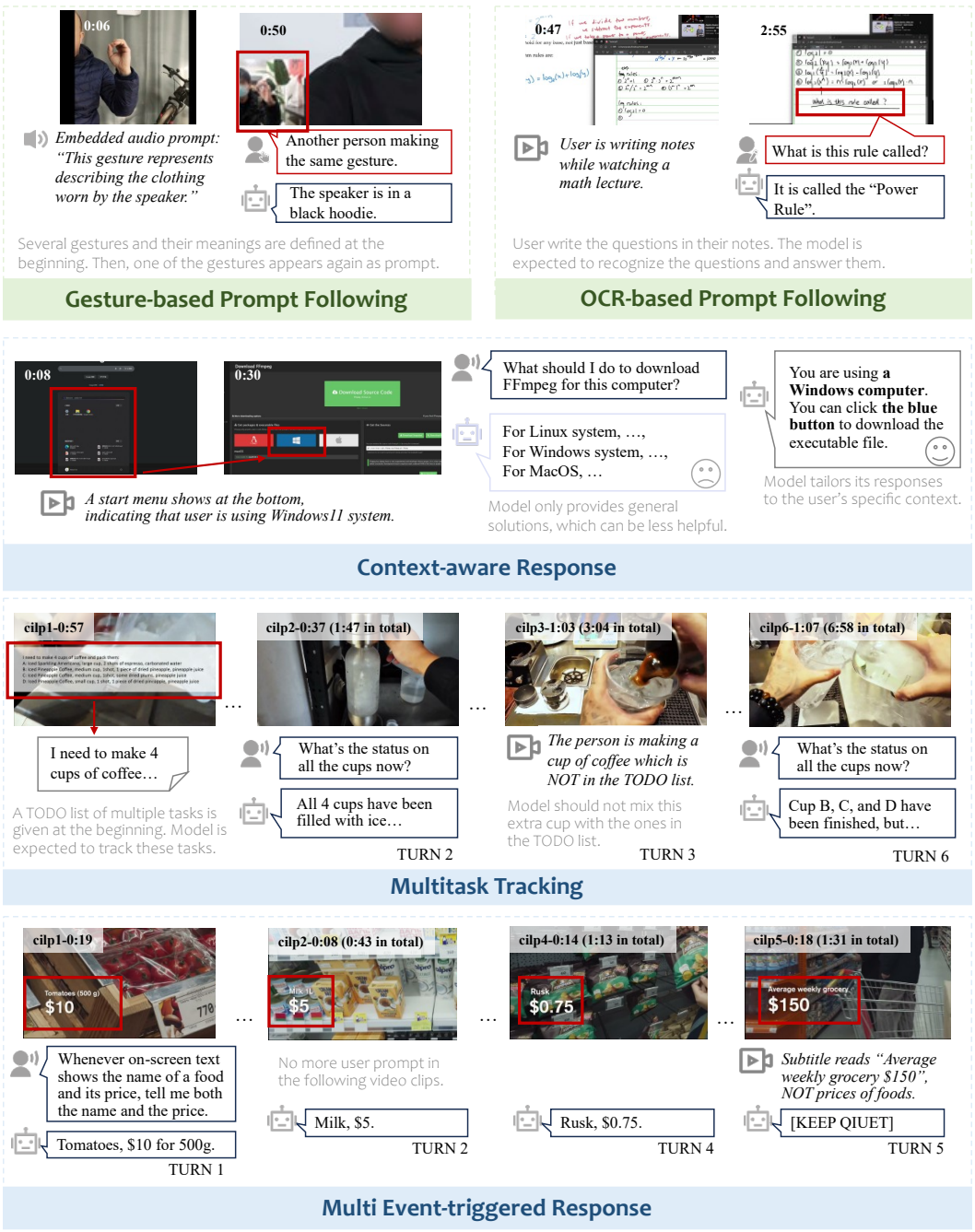


Figure 3: Examples of 5 typical tasks in OmniAssistBench. Gesture-based Prompt Following (GPF) and OCR-based Prompt Following (GPF) are Basic Tier tasks, while Context-aware Response (CR), Multitask Tracking (MT), and Multi Event-triggered Response (MER) are Advanced Tier tasks. Descriptions of the input videos are denoted as *Italic*.

requirement, annotators need to carefully analyze each video, and decide when to ask what so that the interaction plots can be logically correct.

Video editing & user prompt embedding: To simulate realistic interaction, all questions are converted to speech via TTS, volume-balanced, and temporally embedded into the video stream.



Figure 4: Examples of key plots and questions of the 3 Real World Cases. Plot descriptions are denoted as *italic*. *Meeting Simulation* simulates a user watching a round-table discussion involving 12 participants while taking notes. *Blind Assistance* simulates a blind user finding certain object, navigating to a certain location and waiting for a certain event with the help of the assistant. *Handicraft Process Tracking* simulates three individuals making handicraft together while interacting with the assistant, requiring the assistant to track the action of each person independently.

Visual prompts (i.e., gestures and handwriting) are embedded as video clips in a picture-in-picture manner. User goals are presented as on-screen subtitles. Additionally, to accommodate the context window constraints of current MLLMs while preserving task integrity, we manually trim videos for non-temporal tasks to a duration of 30s–180s, ensuring all instructional cues remain intact. All videos are standardized to around 1080p resolution.

Quality Refinement: We employ state-of-the-art MLLMs (e.g., Gemini-3-Pro) to assess annotation completeness. If a model generates a response that is logically correct yet diverges significantly in phrasing from our initial ground truth, we manually verify and append the output as an alternative reference to improve evaluation robustness. However, this augmentation is applied sparingly, as our benchmark's design prioritizes questions with a single, definitive ground truth.

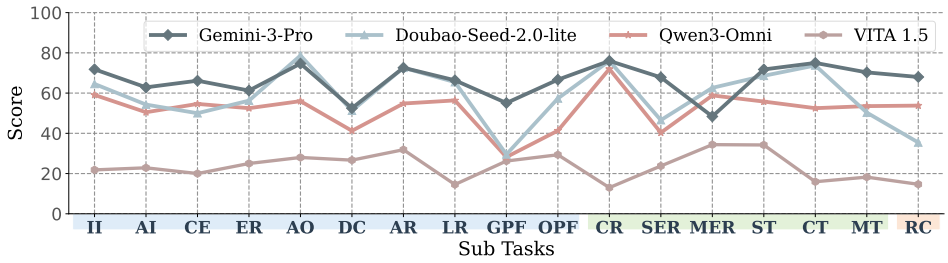


Figure 5: Evaluation results comparison of 4 representative models at sub-task level. Abbreviations of task names are defined in Table 1. The blue, green, and orange backgrounds of the x-axis labels denote basic tier, advanced tier, and real-world tasks, respectively. All original 0-5 scores have been normalized to 0-100.

3 Task Design

Tasks in OmniAssistBench are grouped into the Basic and the Advanced Tier. While the former focus on fundamental perception tasksthat are critical and frequently encountered in real world interactions but are under-represented in existing video benchmarks, the latter valuate higher-level capabilities specific to the model functioning as a interactive assistant. Base on this classification, we distill key abilities and typical scenarios in video assistant interactions to formulate 7 major tasks with 16 sub-tasks. Furthermore, we capture three real world cases. Each case has its plots carefully designed to combine multiple tasks listed in the benchmark and complex scenarios where large-scale data construction is impractical. Typical sub-tasks are demonstrated in Fig. 3. Key plots of the real world cases are demonstrated in Fig. 4. Table 1 summarizes the statistics of the dataset. Detailed explanations of the tasks are provided in Appendix B.

4 Experiments

4.1 Experiment Settings

Evaluated Models: OmniAssisBench requires candidate models to be capable of processing concurrent video and audio streams. Our evaluation suite comprises four closed-source models, **Gemini-2.5-Pro** (Comanici et al., 2025), **Gemini-3-Pro** (Google DeepMind, 2025), **Doubao-Seed-2.0-lite**, **MiMo-V2-Omni** (Xiaomi, 2026), and **Qwen3.5-Omni-Plus** (Qwen Team, 2026), alongside five open-source models: **Qwen3-Omni-Instruct (30B)** (Xu et al., 2025b), **Qwen2.5-Omni (7B)** (Xu et al., 2025a), **MiniCPM-o-4.5** (Yu et al., 2026), **MiniCPM-o-2.6** (Yao et al., 2024), **Baichuan-Omni-1.5** (Li et al., 2025b), and **VITA-1.5** (Fu et al., 2025b). Regarding input preprocessing, all videos are maintained at their native resolution ($\approx$1080p). For models except the Gemini family, videos are sampled at a rate of 1 frame per second (FPS). All other inference parameters adhere to the official recommended configurations respectively.

Context Management: To accommodate the limited context windows of models during long video inputs, we employ a first-in-first-out strategy, discarding the oldest interaction history when the total input token length exceeds the context limit.

Prompting Strategy: We use a unified system prompt to configure the model as a video assistant with proactive abstention, where it autonomously evaluates if visual inputs necessitate a response:

```
You are a helpful video assistant. For the above input video(s), if you decide to output,
organize your output as if you are directly talking to the user. Otherwise, if you decide to
keep quiet, output exactly "[KEEP QUIET]".
```

Scoring Rubric. We evaluate models using a GPT-5 judge on a 5-point rubric, normalized to a 100-point scale. The scores are defined as follows: ≥ 60 represents a valid, at least partially meaningful response; 40 denotes correct instruction comprehension but factual errors; and ≤ 20 reflects a fundamental failure to understand the instruction. For proactive-response tasks, < 60 indicates a failure to respond at the correct time. Detailed rubrics are in Appendix A.

Table 3: Performance comparison on Basic Tier tasks. *Overall Avg.* denotes the average score on the overall benchmark. Best scores are marked in bold. Abbreviations of task names are defined in Table 1. All original 0-5 scores have been normalized to 0-100.

Model	Overall Avg.	Basic Interactive Understanding Tasks														Basic Avg.
		Social Perception				Temporal Perception				Referential Perception			Non-audio Prompt			
		II	AI	CE	Avg.	ER	AO	DC	Avg.	AR	LR	Avg.	GPF	OPF	Avg.	
Proprietary Models																
Gemini-3-Pro	**66.4**	71.8	62.8	66.2	67.0	61.2	74.6	52.6	62.6	72.6	66.4	69.8	55.0	66.6	**57.4**	63.6
Gemini-2.5-Pro	64.6	77.2	59.0	69.6	**68.8**	62.6	72.0	55.0	**63.0**	74.0	78.2	**76.0**	54.8	65.4	56.8	**65.4**
Doubao-Seed-2.0-lite	57.3	64.5	54.3	50.0	55.9	56.3	78.7	51.3	61.7	72.6	65.5	69.4	29.5	57.3	35.0	53.2
MiMo-V2-Omni	53.8	66.6	56.0	64.0	62.4	50.0	77.4	60.0	62.2	67.6	56.4	62.4	27.2	62.6	34.6	53.6
Qwen3.5-Omni-Plus	51.6	61.8	33.4	55.2	50.6	25.0	65.4	50.0	46.4	51.2	50.0	50.6	21.0	37.4	24.4	41.6
Open-source Models																
Qwen3-Omni-Inst.(30B-A3B)	**51.2**	59.0	50.4	54.6	54.8	52.6	56.0	41.2	**49.8**	54.8	56.4	55.6	28.2	41.4	30.8	**46.4**
Qwen2.5-Omni (7B)	43.2	29.0	42.8	48.4	40.6	48.8	52.0	42.6	47.6	48.8	40.0	44.8	21.6	24.0	22.2	37.0
MiniCPM-o-4.5 (9B)	46.0	62.8	53.4	50.8	**55.4**	37.6	65.4	43.8	48.4	60.0	52.8	**56.8**	24.6	24.0	24.4	44.6
MiniCPM-o-2.6 (8B)	40.2	54.6	46.6	44.6	48.4	35.0	46.6	38.8	40.0	47.4	51.0	49.0	33.8	25.4	32.2	41.8
VITA-1.5 (7B)	24.6	21.8	22.8	20.0	21.4	25.0	28.0	26.6	26.6	31.8	14.6	24.0	26.2	29.4	26.8	24.6
Baichuan-Omni-1.5 (7B)	43.2	43.6	40.0	47.0	43.8	43.8	49.4	40.0	44.2	50.4	44.6	47.8	43.2	52.0	**45.0**	45.0

Table 4: Performance comparison on Advanced Tier tasks and Real World Cases. *Overall Avg.* denotes the average score on the overall benchmark. Best scores are marked in bold. Abbreviations of task names are defined in Table 1. All original 0-5 scores have been normalized to 0-100.

Model	Overall Avg.	Advanced Interactive Understanding Tasks								Real World Cases			
		CR	Proactive Response			Process Tracking				Avg.			
			SER	MER	Avg.	ST	CT	MT	Avg.		Ms	Ht	Ba	Avg.
Proprietary Models														
Gemini-3-Pro	**66.4**	76.0	67.8	48.4	58.6	71.8	75.0	70.2	**72.2**	**68.2**	76.4	65.8	65.0	**68.0**
Gemini-2.5-Pro	64.6	74.0	61.2	47.8	54.8	67.8	82.2	69.4	72.0	66.4	51.0	32.8	50.0	44.8
Doubao-Seed-2.0-lite	57.3	76.0	46.6	62.6	54.9	68.6	73.8	50.3	64.9	62.1	26.0	21.4	50.0	35.5
MiMo-V2-Omni	53.8	**78.0**	39.4	46.6	42.8	56.8	61.6	63.0	59.8	55.2	52.6	27.2	44.4	41.0
Qwen3.5-Omni-Plus	51.6	56.0	58.0	63.8	**61.0**	54.6	62.6	53.0	56.2	57.8	51.0	41.6	56.0	50.6
Open-source Models														
Qwen3-Omni-Inst.(30B-A3B)	**51.2**	**72.0**	40.4	58.8	50.0	55.8	52.6	53.6	**54.4**	**53.8**	67.2	40.0	56.0	**53.8**
Qwen2.5-Omni (7B)	43.2	45.0	47.2	55.4	51.2	51.0	32.8	44.4	44.4	46.6	18.2	60.0	49.0	45.2
MiniCPM-o-4.5 (9B)	46.0	63.0	53.2	53.0	**53.0**	44.6	46.0	40.2	43.8	47.8	5.4	50.0	47.0	37.8
MiniCPM-o-2.6 (8B)	40.2	39.0	42.0	32.6	37.4	43.2	43.8	37.6	41.8	40.2	9.0	34.2	41.0	31.2
VITA-1.5 (7B)	24.6	13.0	23.8	34.4	28.8	34.2	16.0	18.2	25.0	25.8	11.0	15.8	16.0	14.6
Baichuan-Omni-1.5 (7B)	43.2	35.0	56.8	48.4	52.8	32.0	40.4	40.6	36.6	41.8	38.2	51.4	44.0	44.8

4.2 Evaluation Results

Tables 3 and 4 present the overall performance of all evaluated models on the Basic and Advanced Tier tasks, respectively. The leading closed-source model, **Gemini-3-Pro**, achieves a score of **66.4**, while the top-performing open-source model, **Qwen3-Omni-Instruct**, scores **51.2**. These results suggest that current MLLMs generally succeed in understanding verbal prompts but struggle with providing accurate and comprehensive answers.

For the Basic Tier tasks: The primary challenges stem from viual perceiving. Models usually struggle to accurately perceive non-salient objects and distinguishing target objects from visually similar distractors. Moreover, even a modest increase in the number of target objects (e.g., more than 10) results in a decline in performance, indicating limited scalability in complex visual scenes.

Models also struggle with following gesture instructions. There might be two reasons. First, from a perception perspective, gestures are inherently abstract and susceptible to visual ambiguity, making them difficult to distinguish from incidental character movements or background visual noise. Second, from a data perspective, current models likely suffer from insufficient exposure to gesture-grounded instructions during training. While gesture prompts are less prevalent than audio commands, they are indispensable for inclusive interaction, particularly for individuals with speech impairments. This underscores an urgent need to address the gap in interpreting subtle, non-verbal visual cues.

For the Advanced Tier tasks: One primary challenge comes from limitation of model context length. With a 32k context window, MiniCPM-o-2.6 can retain approximately 380 seconds of video at 1 fps. Beyond this limit, the model starts to produce nonsensical words. This context capacities are insufficient to preserve a complete interaction history for the multi-turn tasks in our benchmark.

Without specialized mechanisms for long-term memory, context length limitations are likely to become a critical bottleneck for current models.

Another common failure is that models struggle to delay responses for the correct time, neither they state that the current information is insufficient for answering. In turns where no user prompts are present, models are easily distracted by the speeches in the video and forget the original goal.

For the Real World Cases: Evaluated models exhibit several common shortcomings. In the *Handicraft Process Tracking* task, models struggle to track small objects during long-term interaction, especially when moved out of the camera's view. Additionally, models sometimes fail to distinguish whether the utterances are directed at them or other people. For the *Blind Assistance* task, most models fail to accurately understand 2D maps and associate them with 3D real-world environments. Furthermore, in the *Meeting Simulation* task, models generally fail to sustain long-term proactive tasks (i.e., the "whenever xxx then do xxx" tasks). Upon receiving new commands, they tend to switch tasks and struggle to resume the previous "whenever-then" tracking task. These results indicate current models need substantial improvement to become satisfactory assistants.

4.3 Ablation Studies

We conduct four ablation studies whose details can be found in Appendix C. The *input modalities* ablation shows that the model performs worse in most cases with only visual or audio input, indicating that both modalities are necessary. The *judge model consistency* experiments demonstrate that different judge LLMs score the same candidate similarly, suggesting that the final scores are less dependent on the judge models. The *teacher forcing* ablation investigates difficulty balancing for multi-turn questions, while the *low resolution* experiment explores the trade-off of using lower-resolution frames to accommodate longer videos within the context limits.

5 Related Work

Recent advancements have evolved large language models into dynamic Omni-MLLMs capable of real-time, cross-modal reasoning over video, audio, and text streams (Liu et al., 2023; Zhu et al., 2024) into versatile, dynamic Omni-MLLMs (Li et al., 2025a; Wang et al., 2025a; Cao et al., 2025; Li et al., 2024b; Huang et al., 2026; Yin et al., 2026; Wu et al., 2024; Driess et al., 2023; Fei et al., 2025; Xu et al., 2025b). To evaluate these capabilities, existing video understanding benchmarks typically focus on isolated perceptual skills (e.g., streaming, omni-modal, or egocentric perception) mostly using single-turn, textual multiple-choice formats (Mangalam et al., 2023; Li et al., 2024a; Lin et al., 2026; Li et al., 2025c), which fail to reflect real-world video chat scenarios. Compared to standard video understanding, interactive evaluation remains relatively under-explored. For instance, ProactiveVideoQA (Wang et al., 2025b) focuses exclusively on proactive abilities, OmniMMI (Wang et al., 2025c) tends to conflate interactive tasks with basic perception, and EgoLife (Yang et al., 2025) still relies on text-based single-turn questions. Although works like LifeEval (Gao et al., 2026) explore aligning visual inputs with user intentions, they are largely limited to short and visual-only videos. Consequently, a comprehensive, multi-turn omni-modal evaluation pipeline closely reflecting daily interactive usage has yet to be established.

6 Conclusion

This paper introduces OmniAssistBench, a comprehensive benchmark for evaluating Omni-LLMs as interactive assistants. Under interactive scenarios, the model's output directly influence the action of the user, resulting in divergence of interaction paths which static evaluation dataset cannot adapt to. We propose an annotation method to address this problem. We summarize the interaction path based on the source video as prior knowledge. After fixing the path, we are able to reverse-engineering the source video to design test questions and simulate test data by video editing. Based on this annotation pipeline, we build OmniAssistBench and test popular Omni-LLMs. Experiment results indicate the limitations of current models, particularly in gesture instruction following, long-context memory, capacity of delaying response and maintaining information cross turns.

AI USE STATEMENT

In this work, we used generative AI tools for translating, grammar checking, and polishing the English text of the manuscript. Additionally, we used generative AI tools for translation during the dataset construction process. We have not used generative AI tools for directly generating data annotations, devising research methodology, or writing experimental code, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. Specifically, all AI-assisted translations for the dataset were manually checked, verified, and further edited by human annotators prior to inclusion to ensure absolute accuracy and quality. The AI-polished manuscript text was also fully reviewed by the authors to ensure it accurately reflects our original scientific intent. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

REPRODUCIBILITY STATEMENT

To ensure reproducibility, the dataset and example evaluation codes associated with this work have already been fully open-sourced. However, to strictly adhere to the ICLR double-blind review policy, we have temporarily omitted the direct URLs to our project page and public repositories (i.e., GitHub and HuggingFace) in this manuscript, as they contain identifiable author information. We firmly commit to including all exact public links in the camera-ready version upon acceptance.

REFERENCES

Shidong Cao, Zhonghan Zhao, Shengyu Hao, Wenhao Chai, Jenq-Neng Hwang, Hongwei Wang, and Gaoang Wang. Efficient transfer from image-based large multimodal models to video tasks. *IEEE Trans. Multimedia*, 27:3045–3056, 2025.

Gheorghe Comanici, Eric J. Bieber, Mike Schaekermann, Ice Pasupat, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. arXiv:2507.06261. doi: 10.48550/arXiv.2507.06261.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *Proc. ICML*, 2023.

Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, et al. On path to multimodal generalist: General-level and general-bench. In *Proc. ICML*, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proc. CVPR*, 2025a.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yifan Zhang, Yunhang Shen, et al. VITA-1.5: Towards GPT-4o level real-time vision and speech interaction. In *Proc. NeurIPS*, 2025b.

Hengjian Gao, Kaiwei Zhang, Shibo Wang, Mingjie Chen, Qihang Cao, Xianfeng Wang, Yucheng Zhu, Xiongkuo Min, Wei Sun, Dandan Zhu, and Guangtao Zhai. LifeEval: A multimodal benchmark for assistive AI in egocentric daily life tasks. In *Proc. CVPR*, 2026.

Google DeepMind. Gemini 3 pro - model card, 2025. URL `https://blog.google/products-and-platforms/products/gemini/gemini-3/`. Accessed: Aug. 14, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proc. CVPR*, 2022.

Jing Huang, Luyuan Chen, Anpeng Wu, Junkun Yuan, Kun Kuang, Fei Wu, and Qiang Zhu. Rewind-augmented generation for long video understanding. *IEEE Trans. Multimedia*, 2026.

Guozhang Li, Xinpeng Ding, De Cheng, Jie Li, Nannan Wang, and Xinbo Gao. ETC: Temporal boundary expand then clarify for weakly supervised video grounding with multimodal large language model. *IEEE Trans. Multimedia*, 27:1772–1782, 2025a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proc. CVPR*, 2024a.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, et al. Baichuan-Omni-1.5 technical report, 2025b. arXiv:2501.15368. doi: 10.48550/arXiv.2501.15368.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, et al. OmniBench: Towards the future of universal omni-language models. In *Proc. NeurIPS*, 2025c.

Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, Yong Xu, and Min Zhang. LMEye: An interactive perception network for large language models. *IEEE Trans. Multimedia*, 26:10952–10964, 2024b.

Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the gap for MLLMs to achieve streaming video understanding. In *Proc. ICASSP*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proc. NeurIPS*, volume 36, 2023.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Proc. NeurIPS*, volume 36, 2023.

Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, Jake K Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *Proc. CVPR*, 2011.

Yulu Pan, Ce Zhang, and Gedas Bertasius. BASKET: A large-scale video dataset for fine-grained skill estimation. In *Proc. CVPR*, 2025.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proc. ACL*, 2019.

Qwen Team. Qwen3.5-Omni technical report, 2026. arXiv:2604.15804. doi: 10.48550/arXiv.2604.15804.

Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. A multimodal corpus for emotion recognition in sarcasm. In *Proc. LREC*, 2022.

Luca Rossetto, Werner Bailer, Duc-Tien Dang-Nguyen, Graham Healy, Björn ór Jónsson, Onanong Kongmeesub, Hoang-Bao Le, Stevan Rudinac, Klaus Schöffmann, Florian Spiess, et al. The CASTLE 2024 dataset: Advancing the art of multimodal understanding. In *Proc. ACM MM*, 2025.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proc. CVPR*, 2022.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. COIN: A large-scale dataset for comprehensive instructional video analysis. In *Proc. CVPR*, 2019.

Binglu Wang, Yao Tian, Shunzhou Wang, and Le Yang. Multimodal large models are effective action anticipators. *IEEE Trans. Multimedia*, 27:2949–2960, 2025a.

Yueqian Wang, Xiaojun Meng, Yifan Wang, Huishuai Zhang, and Dongyan Zhao. ProactiveVideoQA: A comprehensive benchmark evaluating proactive interactions in video large language models, 2025b. arXiv:2507.09313. doi: 10.48550/arXiv.2507.09313.

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 *Proc. CVPR*, 2025c.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In *Proc. ICML*, 2024.

Xiaomi. Xiaomi MiMo-V2-Omni: Omni-modal agentic foundation model that sees, understands and acts, 2026. URL `https://mimo.xiaomi.com/mimo-v2-omni`. Accessed: Aug. 14, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, et al. Qwen2.5-Omni technical report, 2025a. arXiv:2503.20215. doi: 10.48550/arXiv.2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, et al. Qwen3-Omni technical report, 2025b. arXiv:2509.17765. doi: 10.48550/arXiv.2509.17765.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. EgoLife: Towards egocentric life assistant. In *Proc. CVPR*, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A GPT-4V level MLLM on your phone, 2024. arXiv:2408.01800. doi: 10.48550/arXiv.2408.01800.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Chunjiang Ge, Yan Yang, Yuhan Dai, Yongdong Luo, Tong Xu, Caifeng Shan, and Enhong Chen. Sparrow: Data-efficient Video-LLM with text-to-image augmentation. *IEEE Trans. Multimedia*, 2026.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, et al. MiniCPM-V 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proc. CVPR*, 2026.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *Proc. ICLR*, 2024.

A SCORING RUBRIC

Table 5 demonstrates the detailed scoring rubric provided to the judge models.

Table 5: The Scoring Rubric used in OmniAssistBench. The evaluation is based on three dimensions: Accuracy and Completeness, Redundancy, and Adherence to Silence Constraints. These rules are provided to the judgment LLM as prompts during evaluation.

Score	Accuracy & Completeness	Redundancy	Response to Silence *(GT = "KEEP QUIET")*
5	**Perfect Execution.** The model correctly comprehends the intent and semantically matches the Ground Truth (GT). It covers *all* Key Points.	**None.** Contains almost no redundant information; only minor, contextually necessary elaborations are allowed.	**Exact Match.** Correctly judges that silence is needed and outputs the exact required token "[KEEP QUIET]".
4	**Minor Defects.** Comprehends the intent and covers all Key Points, but the response contains minor factual errors or lacks precision in description.	**Low.** Contains minor redundant information, provided it remains relevant to the core subject.	**Descriptive.** Describes the state of inaction but fails to follow the specific instruction to output the exact silence token.
3	**Partial Success.** Comprehends the intent but misses at least one Key Point. The answer is directionally correct but factually imperfect.	**Moderate.** Contains information not present in the GT, but it remains relevant to the general context of the video.	**Evasive.** States that there is insufficient information to answer.
2	**Contradiction.** Comprehends the intent but provides a clearly wrong answer (e.g., opposite polarity, wrong object) or directly contradicts the GT.	**High.** Contains significant irrelevant information, obvious hallucinations, or details unrelated to the query.	**Hallucinated Response.** Fails to understand that current turn is not a proper time for response. Gives hallucinated answers to the questions instead.
1	**Loose Relation.** Fails to comprehend the specific question. The response is only loosely related to the topic.	—	**Loose Relation.** Provides a response that is tangentially related to the context but completely ignores the silence constraint.
0	**Failure.** Complete failure to comprehend the input. The response is entirely unrelated to the question, instruction, or video content.	—	**Failure.** Output is entirely unrelated to the question or context.

B Detailed Task Lists

Table 6 and Table 7 list all tasks in OmniAssistBench. The italic words briefly explain the key properties which are used for filtering videos or designing QAs during the data collection process.

B.1 Basic Interactive Understanding

Regarding Basic Interactive Understanding capabilities, we focus on fundamental perception tasks that are critical and frequently encountered in real world interactions but are under-represented in existing video benchmarks. Below are introductions of the major tasks in this category. Table 6 lists all the sub tasks contained in this tier.

Social Perception. Assistant are excepted to recognize user identities, addressees, and emotions. We use longer videos to test social perception abilities based on complete scenes, instead of only one or two utterances.

Temporal Perception. Users may expect the assistant to help them recall overlooked details in the past. Motivated by this need, we focus on evaluating the perception on non-salient details instead of the main objects in the video.

Referential Perception. User may refer to objects by describing their locations or pointing at them. We design challenging cases where model need to distinguish the exact object referred by the user among multiple similar ones.

Non-audio Prompt Following. Beyond verbal commands, users may also convey instructions through hand gestures or text written on physical objects. These visual modal instructions are under-explored in current benchmarks.

B.2 Advanced Interactive Understanding

This tier evaluate higher-level capabilities specific to the model functioning as a interactive assistant, derived from typical application scenarios. Table 7 lists all the sub tasks contained in this tier.

Context-Aware Response. To be effective, assistants need to go beyond generic suggestions and offer actionable advice grounded in the user's specific environment. For example, model may suggest that user can use the tool right next to their hand, instead of only suggest using tools of that kind.

Proactive Response. In this setting, the target event occurs some time later after the user's query. Models must determine not only *what* to answer but also *when* to answer.

Process Tracking. This task assesses the model's ability to provide continuous assistance as the user performs long-horizon, multi-step tasks.

B.3 Real World Cases

Real world interaction scenarios are usually far more complicated than the single tasks designed above. We choose 3 representative real world application scenarios of assistant models, and carefully design the plots so that multiple abilities can be evaluated in the same story. We then film one-shot videos base on the scripts, resulting in 3 test cases. Key plots are demonstrated in Fig. 4.

Meeting Simulation. This scenario simulates a user participating in a round-table discussion involving 12 participants. The model need to process a continuous 20-minute video stream to support the user in note-taking and information retrieval. Key challenges include fine-grained speaker identification, understanding hand gestures (understand when user tasks screen shots), and maintaining long-term memory to recall details scattered throughout the interaction.

Blind Assistance. Visually impaired individuals may depend on assistant models for environment description and guidance. While constructing large-scale benchmarks for such tasks is challenging due to environmental complexity, we design a high-fidelity case study to rigorously evaluate this capability. In this scenario, a user simulates visual impairment and requests assistance for three tasks in an office setting: visually grounding specific objects, navigating to a destination via a hand-drawn map, and asking for a reminder when a certain person comes. Apart from navigation, this task also test model's proactive response ability.

Handicraft Process Tracking. This case is a combination of multi-party social interaction and process tracking. In the video, three individuals are making handicraft together. Model need to recognize all the people and distinguish whether an utterance is directed at the assistant or at another human participant. Also, model need to constantly track small handicrafts even if they are temporally moved out of the camera's view.

Table 6: Detailed table of all Basic Tier sub-tasks. *Italic* sentences are explanations of the task.

Major Task	Abbr.	Sub Task	Description	Turns
Social Perception	**II**	**Identity Identification**	"Who said that sentence? Who did that action?" *Models need to associate multiple actions or speeches with the same person, instead of simply giving answers like "the man in red" which can be derived from visual information directly without understanding actual identity.*	**Single turn**
	AI	**Addressee Identification**	"Who are they talking to?" *There are always multiple people in the video apart from the speaker and the addressee. Models may need to check the responses of the people to understand who is the addressee.*	**Single turn**
	CE	**Complex Emotion Understanding**	"What emotion or attitude is the person expressing?" *We only keep complex samples where one modality (e.g., expression or dialogue) alone cannot lead to the correct answer.*	**Single turn**
Temporal Perception	**ER**	**Event Retrieval**	"When the main person A said xxx, what color clothes were the person passing behind them wearing?" *The model needs to establish temporal correlations between the main and non-salient objects or details.*	**Single turn**
	AO	**Appearance Order Perception**	"Which colors of clothes did the blogger try on in sequence?" *Model may need to distinguish the target object from similar ones (e.g., for example, the blogger may have picked up a lot of clothes but only tried on a part of them).*	**Single turn**
	DC	**Dynamic Counting**	"How many standard pull-ups did the person perform?" *Models need to count a large number of instances (> 10) or filter for specific objects among similar ones.*	**Single turn**
Referential Perception	**AR**	**Action Reference Perception**	"Which student is the teacher pointing at? Which file is the mouse point at on the screen?" *There are always multiple objects similar to the target. In some cases, the referred object cannot be decided by only looking at the "pointing" action. Models may need to look for other cues (e.g., there may be multiple students in the direction where the teacher is pointing at, but only the target student stand up after being pointed).*	**Single turn**
	LR	**Linguistic Reference Perception**	"What is the title of the second book on the right?" *There are multiple objects similar to the target. The target object may not be the main object of the video.*	**Single turn**
Non-audio Prompt Following	**GPF**	**Gesture-based Prompt Following**	"This gesture represents describing the cloth wore by the speaker." *4 gestures are defined at the beginning of the video, then these gestures are embedded using a picture-in-picture layout as prompts. The gestures as prompts may be made by a different person. Not all pre-defined gestures necessarily appear as prompts, nor do they strictly follow the defined order*	**4 turns in avg.**
	OPF	**OCR-based Prompt Following**	"What is the name of this calculation rule?" *User prompts are embedded as on-screen subtitles or video clips of hand writing using a picture-in-picture layout. There are always other English characters in the video apart from the prompt.*	**Single turn**

Table 7: Detailed table of all Advanced Tier sub-tasks. *Italic* sentences are explanations of each task.

Major Task	Abbr.	Sub Task	Description	Turns
Proactive Response	**SER**	**Single Event-triggered Response**	"When the taxi comes, tell me its license plate number." *Models need to wait until the target object appear, then give the answer. We do not explicitly ask the model whether to respond at the end of each turn. Instead, the user prompt is provided only in the first turn, followed by raw video clips in subsequent turns. The model must autonomously determine the appropriate timing for a response.*	**4 turns in avg.**
	MER	**Multi Event-triggered Response**	"Whenever the professor flips to a new slide, tell me its title." *Models need to wait until the target object appear. The target event may appear several times, but the user prompt appears only once in the first turn.*	**4 turns in avg.**
Process Tracking	**ST**	**Step Tracking**	"I want to make this dish strictly following this example video. What should I do next?" *To avoid ambiguity, we provide a reference (manual images or demonstration videos) at the beginning. User might ask for the next step after completing multiple steps or while halfway through one, preventing shortcut responses that merely recite the next sentence*	**6 turns in avg.**
	CT	**Checklist Tracking**	"This is my shopping list. Which items have not been bought yet?" *A checklist is provided at the beginning as an image rendered from text. Models need to remember which objects have appeared. Objects may appear in a different order compared with the list.*	**4 turns in avg.**
	MT	**Multitask Tracking**	"Here is a TODO list and brief steps of how to finish each task. What is the progress of everything now?" *The user constantly switch between multiple tasks. There are strict sequential steps within a task, but no dependency between tasks. Models need to track the status of each task, and remind the user whenever required.*	**4 turns in avg.**
Context-Aware Response	**CR**	**Context-Aware Response**	"I'm on the FFmpeg download page. How can I download FFmpeg for this computer?" *It can be inferred from the video that the user is using a Windows machine. We prefer the model to tailor its suggestions exclusively to the Windows platform, instead of providing generic suggestions covering all major operating systems.*	**Single turn**

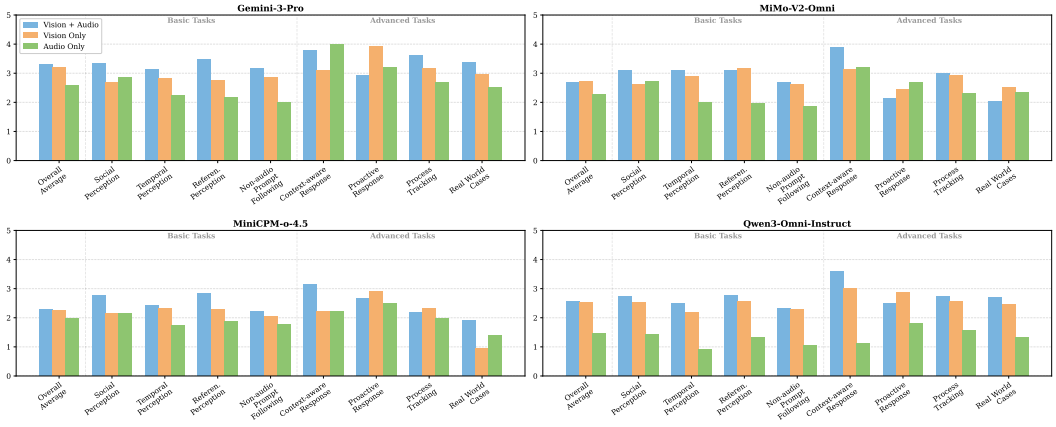


Figure 6: Major-task level performance comparison of four candidate models under original, visual-only, and audio-only input conditions. While removing visual or audio inputs generally degrades performance, models perform better in the Proactive Response task with visual-only inputs, as this prevents them from mistakenly answering background speech instead of the user query. Additionally, Gemini-3 scores higher in the Context-aware Response task under the audio-only setting, as it avoids formatting penalties triggered by outputting timestamps when video is present.

C Detailed Ablation Studies

In this section, we provide detailed analyses of four ablation studies. The *input modalities* ablation shows that the model performs worse in most cases with only visual or audio input, indicating that both modalities are necessary. The *judge model consistency* experiments demonstrate that different judge LLMs score the same candidate similarly, suggesting that the final scores are less dependent on the judge models. The *teacher forcing* ablation investigates difficulty balancing for multi-turn questions, while the *low resolution* experiment explores the trade-off of using lower-resolution frames to accommodate longer videos within the context limits.

C.1 Ablation on Input Modalities

To understand the necessity of the input modalities, we conduct an ablation study by removing either the audio or the visual components from the original test samples. For the audio-only ablation, we simply removed all video frames. For the visual-only ablation, all audios are removed. Consequently, user prompts originally given as speech were converted into text. Prompts given as subtitles remained unchanged. For gesture prompts, the original spoken instructions defining the meaning of the gestures were converted to text, while the gestures themselves remained. Models still need to understand the gestures. We tested four candidate models under these conditions: Gemini-3-Pro, Mimo-V2-Omni, MiniCPM-o-4.5, and Qwen3-Omni-Instruct. Major-task level results are summarized in Fig. 6.

The histogram shows that on most tasks, removing either the visual or the audio modality causes a decrease in performance. This drop is especially noticeable for the audio-only inputs. However, for the visual-only inputs, the decrease in scores was not very obvious for some tasks. A potential reason is that those models already had relatively low scores on the original inputs. Because they were initially struggling to process the combined audio and video, removing the audio mostly caused them to swap one incorrect answer for a different incorrect answer, resulting in little change to their overall score.

Despite the general drop in performance, we observe an interesting phenomenon: on the ”Proactive Response” task, all models actually performed better with visual-only inputs. The proactive response task is a multi-turn task where the user prompt is only given in the first turn. In the following turns, the model receives new multimodal inputs and need to decide on its own whether to respond. In the original, full-audio inputs, models tend to mistook background speech in the video for new instructions from the user, causing them to forget the original prompt. In the visual-only setting, the absence of distracting audio allowed the models to stay focused on the initial user prompt. This

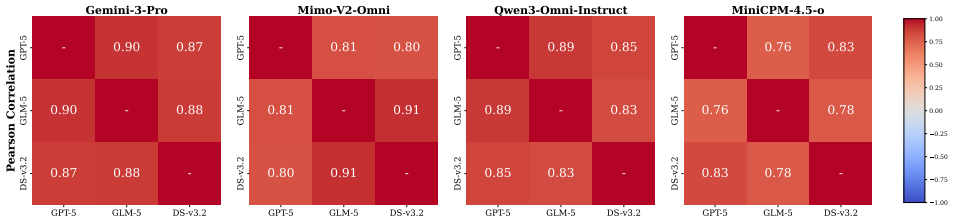


Figure 7: Heat maps of Pearson correlation coefficients between evaluation scores from different judge LLMs across 4 candidate models. DS-v3.2 stands for DeepSeek-v3.2. All coefficient values are above 0.75, indicating that different judge models are likely to agree with each other on whether a candidate's answer is good or not.

suggests a potential limitation in current models: they tend to blindly reply to any spoken query they hear, rather than successfully identifying and keeping track of a specific user across multiple turns.

Also, we notice that Gemini-3-Pro performs better with audio-only input on the Context-aware Response task. According to the model replies, this result can be attribute to the model's output format: With video input, model tend to follow a markdown format of "timestamp–cue description", which is punished by our scoring rubric for being less aligned with spoken language. However, model tend to output colloquial language for audio-only inputs, thus scoring higher.

C.2 Consistency of Different Judge Models

To understand how the choice of the evaluator affects the results, we conducted an ablation study comparing three different judge LLMs: GPT-5, GLM-5, and DeepSeek-v3.2. In this experiment, each judge model is given the exact same scoring criteria in its prompt and asked to evaluate the outputs of four different candidate models (Gemini3-Pro, Mimo-V2-Omni, MiniCPM-o-4.5 and Qwen3-Omni-Instruct). To measure how closely the judge models agreed with one another, we calculated the Pearson correlation coefficient for every pair of judges evaluating the same candidate. As shown in the four heat maps in Fig. 7, all correlation values are close to or higher than 0.8. This strong positive correlation indicates a high level of agreement among the three judge models. It shows that, regardless of the specific judge LLM used, the scores can reliably reflect the actual performance of the candidate models.



Figure 8: Performance comparison on multi-turn tasks under the common setting and the teacher forcing setting, where historical model responses are replaced by ground truth answers to prevent error accumulation and bridge the difficulty gap between multi- and single-turn tasks. The results indicate that independently evaluating each question in a multi-turn interaction does not always improvements model scores. All original 0–5 scores have been normalized to 0–100.

C.3 Teacher Forcing Test Setting for Multi-turn Tasks

Multi-turn questions are usually considered more difficult than single-turn questions, as errors that accumulated in previous turns may influence model's performance on the following turns. To address this question, we tested the open-sources models under a *teacher forcing* setting: When obtaining the model output at the Tth turn, we replace all preceding model responses in the $T-1$ turns of

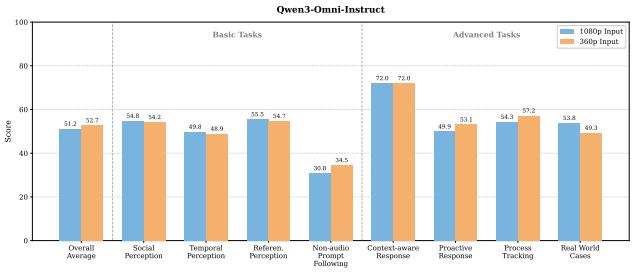


Figure 9: Performance comparison of Qwen3-Omni-Instruct under different input video frame sizes (1080p and 360p). Reducing the input resolution to fit more video frames within the same context length does not significantly improve performance. All original 0–5 scores have been normalized to 0–100.

interaction history with ground truth answers. This prevents error propagation from previous turns, ensuring that the prediction for the current turn is not affected by historical inaccuracies.

The results comparison of the common setting (without changing the chat history) and the teacher forcing setting are illustrated in Fig. 8. Although providing a correct interaction history can improve the scores, the improvement is not always obvious. We observe that models quickly adopt to the format of the ground truth answer, which are usually briefer than common model output. However, models still fail to link the ground truth answer with the input videos and take advantages from them.

C.4 Influence of Input Video Resolution

To address the trad-off between the resolution of input video frames and the total number of frames that can be hold by the maximum context length of the open-source models, we evaluated Qwen3-Omni-Instruct using resized videos where the minimum edge length is fixed to 360 pixels. Under this setting, the model is able to accommodate temporal context of approximately 50 seconds longer than 1080p input.

As shown in Fig. 9, although holding more video frames in the model context slightly increase the scores on most multi-turn questions, there is no obvious change on the scores across the whole benchmark. On one hand, 50 seconds longer context may still be insufficient for long videos that may last more than 10 minutes, such as videos in the Real-World Cases task. On the other hand, context limitation may not be the only bottleneck that prevent the model from giving valid responses.