CAN OMNI-MODAL LLMS UNDERSTAND SPOKEN INTENTIONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal large language models (Omni-modal LLMs) can answer questions about audiovisual content, yet useful assistance also requires understanding what help the user intends to obtain. This paper introduces OMNIINTENT-BENCH for evaluating spoken intention understanding through contextual responses. The central requirement is to determine the valid request, retain its constraints, and respond according to the available evidence. We accordingly organize the benchmark around request grounding, request and context selection, assistance need assessment, request integration, and responding under uncertainty. Beyond direct requests, it examines indirect help seeking, revised requirements, ambiguous goals, and false premises. Paired scenarios vary request and evidence conditions to test whether models adjust their responses appropriately. A scenario-driven pipeline constructs scripts, generates audiovisual interactions, and defines response rubrics for language, intent, and content. Recorded videos with synthesized questions provide complementary spoken QA evaluation and training data. We further develop OMNIINTENT-RL through reinforcement learning to examine transfer from explicit spoken QA to complex request conditions. Evaluation reveals substantial variation across models and tasks. Post-training improves instruct-mode performance, while uncertainty handling and selective responding remain weak. These findings establish OMNIINTENT-BENCH as a diagnostic resource for assessing spoken intention understanding and identifying priorities for model improvement.

1 INTRODUCTION

Omni-modal large language models allow users to seek assistance through speech and video (Hurst et al., 2024; Xu et al., 2025b). Useful assistance requires more than recognizing what is said and shown. A user expressing frustration as a yoga mat rolls back up may be asking for practical help without posing a question. Conversely, an explicit question addressed to a companion may require no response from the assistant. The model must determine whether help is requested, what goal remains active, and what the scene supports. These requirements motivate our central question: *Can omni-modal LLMs understand spoken intentions?* We define a spoken intention as the user's current assistance goal, including its target and constraints as expressed through speech and the scene.

Video understanding benchmarks commonly provide the question as text, taking the requested task as given (Fu et al., 2025a; Hong et al., 2026; Li et al., 2026). Dialogue benchmarks bring spoken requests and user goals into evaluation (Pourreza et al., 2026; Wang et al., 2025; Sun et al., 2026), but leave important request conditions without dedicated diagnostic tasks, as summarized in Tab. 1. An indirect request can be clear from context, while an explicit request may still require clarification. A response may also follow a canceled requirement, overlook another speaker's constraints, or accept a false premise. These cases call for evaluating whether the model identifies the valid assistance goal and chooses a response supported by the available evidence.

We introduce OMNIINTENT-BENCH to examine these conditions. Its diagnostic subset contains 1,400 generated audiovisual clips across five categories and 14 subtasks, illustrated in Fig. 1. Scenario scripts specify the spoken request, observable evidence, and expected response behavior. Paired scenarios vary request ownership, trigger conditions, goal clarity, or available evidence, testing both inappropriate responses and unnecessary withholding of help. A spoken QA subset contains approximately 500 recorded videos with synthesized questions. Models receive a complete clip

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.

and a shared instruction, without a question transcript, and produce a textual reply or `[SILENCE]`. Rubrics assess language, intent alignment, and content while accepting different valid responses. This protocol examines response semantics without measuring response timing.

We further ask whether explicit spoken QA training transfers to complex request conditions. We construct OMNIINTENT-TRAIN-24K from recorded videos and synthesized questions, selecting approximately 24k samples through quality and difficulty filtering. We mix these samples with general multimodal data and train an internal model using Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Training combines instruct and thinking modes with rewards for instruction following and response quality. The resulting OMNIINTENT-RL allows us to examine whether gains extend to behaviors such as clarification and selective responding, which the spoken QA rewards do not explicitly supervise.

Evaluation across twelve model configurations shows that request grounding is consistently stronger than assistance need assessment and uncertainty handling. Gemini3.8-Flash performs best overall, while most evaluated models struggle to remain silent when required. In instruct mode, OMNIINTENT-RL scores 66.6 versus 53.9 for general-data RL, but its thinking-mode score remains comparable and selective responding remains weak. These findings answer our title question with *not yet reliably*: stronger performance on grounded requests does not ensure appropriate responses across request conditions.

Our contributions are fourfold:
• **Problem formulation**. We study spoken intention understanding through responses that respect the current assistance goal, contextual evidence, and valid constraints.
• **Benchmark**. We introduce OMNIINTENT-BENCH with 14 diagnostic subtasks, paired scenario conditions, and a complementary spoken QA subset.
• **Post-training**. We construct OMNIINTENT-TRAIN-24K and develop OMNIINTENT-RL to examine the transfer of explicit spoken QA training to complex request conditions.
• **Empirical diagnosis**. We identify differences across models, request conditions, and output modes, showing that stronger performance can coexist with poor uncertainty handling and selective responding.

2 Related Work

2.1 Multimodal Intent Understanding

Video-MME, WorldSense, and OmniVideoBench (Fu et al., 2025a; Hong et al., 2026; Li et al., 2026) evaluate content understanding through externally supplied multiple-choice questions. This format fixes the question for the model, leaving spoken request identification untested. Streaming benchmarks extend evaluation to understanding as videos unfold (Lin et al., 2026; Niu et al., 2025). QIVD (Pourreza et al., 2026) studies situated spoken questions, including visual reference resolution and answer timing. OmniMMI (Wang et al., 2025) covers temporal states, dialogue dependencies, speaker identification, proactive alerts, and selective responding. OmniAssistBench (Sun et al., 2026) reconstructs assistance sequences from existing videos, using predefined priors to evaluate responses conditioned on user goals and visual states. SocialOmni (Xie et al., 2026) separately diagnoses speaker identification, turn entry, and response generation in multiparty dialogue.

MIntRec2.0 (Zhang et al., 2024) classifies utterance intentions and detects inputs outside predefined categories, while WAGIBench (Veerabadran et al., 2025) infers user goals from egocentric multimodal context. Such intent or goal predictions do not evaluate whether the assistant's response fulfills the inferred request. OmniInteract and VideoFDB (Lu et al., 2026; Mazumdar et al., 2026) further evaluate streaming responses and conversational turn control. However, speaker identification and turn-entry decisions do not establish whether a response integrates multiple valid requests and retains their constraints. The dialogue benchmarks in Tab. 1 do not organize implicit requests, request revisions, false premises, and clarification into dedicated diagnostic tasks.

OMNIINTENT-BENCH examines these conditions through responses to spoken requests, including the integration of multiple speakers' goals and constraints. It tests whether models seek clarification or missing evidence, correct false premises, and remain silent when no response is required. Paired scenarios vary request or evidence conditions to expose both inappropriate responses and unnecessary

Table 1: **Spoken intention evaluation across multimodal benchmarks.** We select representative benchmarks for video understanding, situated dialogue and assistance, and streaming or full-duplex conversation. The comparison covers spoken input, complex requests, response behaviors, and paired scenarios. The diagnostic subset of OMNIINTENT-BENCH emphasizes interpreting valid requests, retaining their constraints, and choosing appropriate responses as request conditions change.

Benchmark	Input	Request types					Response behavior		Data design
	Spoken requests	Implicit requests	Multi-speaker requests	Request revisions	Multiple constraints	False premises	Clarification	Silence	Paired scenarios
General video understanding									
Video-MME (Fu et al., 2025a) CVPR'25	✗	✗	✗	✗	✗	✗	✗	✗	✗
WorldSense (Hong et al., 2026) ICLR'26	✗	✗	✗	✗	✗	✗	✗	✗	✗
OmniVideoBench (Li et al., 2026) ICLR'26	✗	✗	✗	✗	✗	✗	✗	✗	✗
Situated dialogue and assistance									
QIVD (Pourreza et al., 2026) ICLR'26	✓	✗	✗	✗	✗	✗	✗	✓	✗
OmniMMI (Wang et al., 2025) CVPR'25	✓	✗	✗	✗	✗	✗	✗	✓	✗
OmniAssistBench (Sun et al., 2026) arXiv'26	✓	✗	✓	✗	✓	✗	✗	✓	✗
SocialOmni (Xie et al., 2026) arXiv'26	✗	✗	✗	✗	✗	✗	✗	✓	✗
Streaming and full-duplex conversation									
OmniInteract (Lu et al., 2026) arXiv'26	✓	✗	✗	✗	✗	✗	✗	✓	✗
VideoFDB (Mazumdar et al., 2026) arXiv'26	✓	✗	✗	✗	✗	✗	✗	✓	✗
Our benchmark									
OMNIINTENT-BENCH	✓	✓	✓	✓	✓	✓	✓	✓	✓

Coverage refers to explicit evaluation targets. Paired scenarios share a scene design and vary request conditions.

withholding of help. Separate intent and content criteria distinguish request interpretation from answer correctness. This semantic diagnosis complements online evaluation through a complete-clip protocol without response timing.

2.2 POST-TRAINING FOR OMNI-MODAL UNDERSTANDING

Models such as Qwen3-Omni (Xu et al., 2025b) and OmniVinci (Ye et al., 2026) support joint audio-visual understanding. GRPO (Shao et al., 2024) provides an efficient framework for reinforcement learning, with DeepSeek-R1 (Guo et al., 2025) demonstrating gains in reasoning. EchoInk-R1 (Xing et al., 2025) applies GRPO with accuracy and format rewards to multiple-choice questions over synchronized audio and images. AVATAR (Kulkarni & Fazli, 2026) combines experience replay with temporal advantage shaping to improve sample efficiency and credit assignment in video reasoning. OmniVideo-R1 (Chen et al., 2026) uses GSPO to optimize query-relevant grounding and modality-attentive fusion, strengthening the use of evidence across modalities. These methods emphasize answer reasoning and evidence use for a given query. We instead study how training on explicit spoken QA transfers to requests with implicit goals, revised constraints, or insufficient evidence. We combine real videos with synthesized requests, filter samples for quality and difficulty, and apply mixed-data GRPO with rewards for instruction following and response quality. The diagnostic benchmark tests the scope of this transfer beyond aggregate QA accuracy, including behaviors not explicitly supervised by the training data.

3 DATASET FOR OMNIINTENT

Constructing OMNIINTENT-BENCH requires making both a user's request and the evidence needed to interpret it observable in a video. A spoken question alone cannot establish whether it addresses the assistant or another person. We therefore first define the request conditions to examine, then construct scenes that express them through speech and visible actions. Sec. 3.1 presents the task taxonomy, followed by scenario generation and annotation in Secs. 3.2 and 3.3. Sec. 3.4 describes the complementary OMNIINTENT-TRAIN-24K, built from recorded videos with synthesized questions. Sec. 3.5 summarizes the data resources.

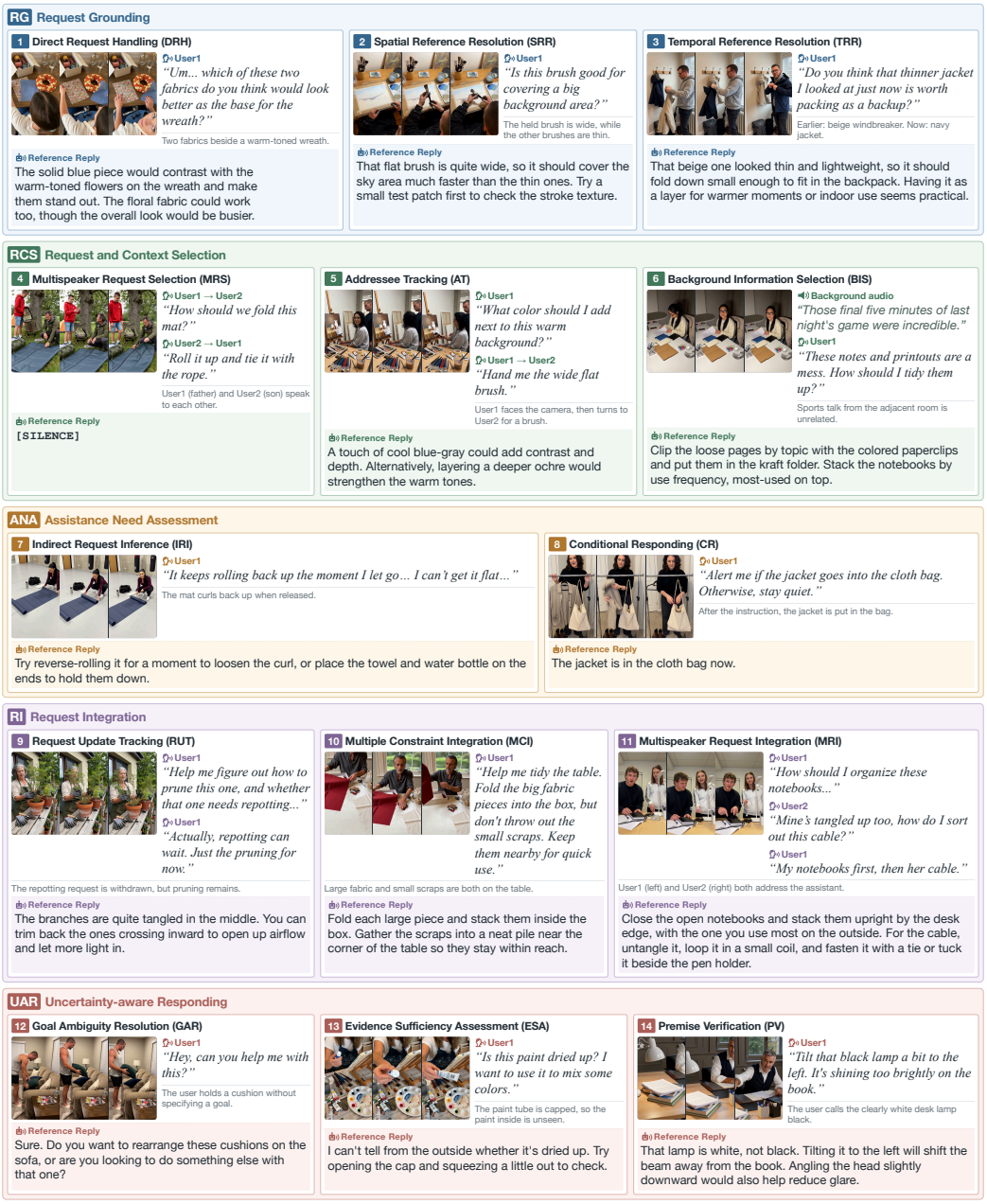


Figure 1: **Overview of the diagnostic subset of `OmniIntent-Bench`.** This subset comprises 14 subtasks organized into five categories of spoken intention understanding. Each panel presents an example with video frames, spoken utterances, relevant scene context, and a reference reply. Audiovisual context guides assistance, clarification, or silence. Reference replies are illustrative and excluded from evaluation.

3.1 Task Definition and Taxonomy

We define a *spoken intention* as the user's current assistance goal, target and constraints conveyed through speech and the scene. Given a complete audiovisual clip, the model provides assistance, clarification or a qualified answer, or returns `[SILENCE]` when no response is required. Response timing and subsequent dialogue turns are outside the scope. A shared instruction guides interaction behavior, without providing sample-specific question transcripts, labels, rubrics or reference replies.

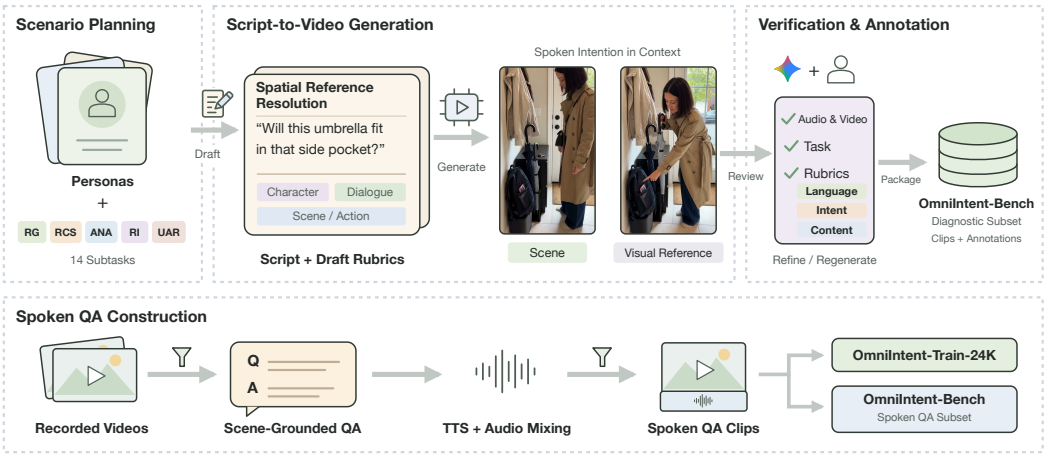


Figure 2: **Data construction pipeline of `OmniIntent`.** The upper route generates the diagnostic subset. The lower route adds synthesized questions to recorded videos to construct the spoken QA subset and OMNIINTENT-TRAIN-24K.

Fig. 1 illustrates the five categories and 14 subtasks with representative scenarios.

Request Grounding (RG). *Direct Request Handling (DRH)* tests explicit requests with sufficient scene evidence. *Spatial Reference Resolution (SRR)* resolves objects or locations through gestures and spatial relations. *Temporal Reference Resolution (TRR)* resolves references to earlier objects, states or events beyond the final frame.

Request and Context Selection (RCS). *Multispeaker Request Selection (MRS)* separates assistant-directed requests from conversations between people. *Addressee Tracking (AT)* follows a speaker's changes of addressee while retaining uncanceled assistant requests. *Background Information Selection (BIS)* uses relevant background speech as evidence and excludes unrelated content.

Assistance Need Assessment (ANA). *Indirect Request Inference (IRI)* recognizes help sought through an expressed difficulty and its audiovisual context. *Conditional Responding (CR)* follows an explicit agreement to respond when a specified event occurs and remain silent otherwise.

Request Integration (RI). *Request Update Tracking (RUT)* incorporates corrections and cancellations, excluding superseded requirements. *Multiple Constraint Integration (MCI)* jointly satisfies compatible preferences and restrictions, including negated instructions. *Multispeaker Request Integration (MRI)* combines valid assistant-directed requests from multiple participants.

Uncertainty-aware Responding (UAR). *Goal Ambiguity Resolution (GAR)* addresses unclear goals through targeted clarification or conditional guidance. *Evidence Sufficiency Assessment (ESA)* acknowledges missing information and seeks relevant evidence. *Premise Verification (PV)* corrects premises contradicted by observable evidence. Each includes a normal counterpart to test for unnecessary clarification or correction. Sec. A details task boundaries and condition variants.

3.2 BENCHMARK CONSTRUCTION

Fig. 2 shows two construction routes, generating diagnostic interactions from scripts and adding spoken questions to recorded videos for the spoken QA subset and OMNIINTENT-TRAIN-24K. We combine PersonaHub descriptions (Ge et al., 2024) with the 14 subtasks to design everyday interactions. Each scenario specifies the request, a potential misinterpretation, observable evidence and expected behavior. Personas guide characters and settings without adding hidden response requirements. A reasoning model produces scripts and draft rubrics, followed by joint speech and video generation. Chinese and English versions preserve the request's meaning, while condition variants alter relevant interaction cues. Variants share a scenario design but need not be visually

identical. Reference replies and rubrics remain outside the media. We check media quality, task validity and rubric consistency, revising annotations or regenerating clips as needed. Accepted clips form the diagnostic subset. Sec. 3.3 describes the sample annotations. For recorded videos, we generate scene-grounded QA pairs and mix synthesized questions into the audio. Screened clips supply separate evaluation and training sets, following the construction process in Sec. 3.4.

3.3 Annotation and Data Separation

Annotations describe the requests and evidence in the generated clips.

Sample annotations. Each synthetic sample records its task, language, scenario group, condition, response requirement, dialogue and scene context. Three sample-specific rubrics specify the response language, the valid request and expected behavior, and the necessary facts and constraints, corresponding to *language*, *intent* and *content*. An example response illustrates one acceptable reply and supports annotation review. It is not a unique ground-truth answer and is supplied to neither the model nor the judge.

Data separation. Training data are fully disjoint from both evaluation subsets of OmniIntent-Bench. No evaluation video or associated QA pair is used for post-training.

3.4 Training Data

The lower route in Fig. 2 constructs spoken QA data from recorded videos. A portion is reserved for the spoken QA subset of OmniIntent-Bench, while the remaining pool supplies training data for OmniIntent-Train-24K. Both resources therefore share a construction process, with separate roles in evaluation and post-training. We retrieve candidate videos from WeChat Channels using scene categories and title keywords, including outdoor exploration, animals, cooking and crafts. A multimodal model filters for continuous, clearly visible footage suitable for a video-call setting. The filtering criteria exclude montages, frequent shot changes, prominent overlays, non-camera content and severe quality defects. Sampling across scene categories limits concentration in a few common settings. For each retained video, a multimodal model generates one conversational question and selects an appropriate point for its insertion. Questions cover visual comprehension, such as fine-grained recognition and temporal tracking, and reasoning, such as comparison and causal explanation. The generation instructions require a natural information need grounded in the scene and discourage generic questions answerable without the video or wording that reveals the answer. A separate generation step produces a concise reference answer containing the key facts. Answer generation is instructed to use video evidence available up to the question time, together with relevant source metadata. The reference answer provides factual supervision for post-training. We synthesize each question using its associated voice and tone description and insert the speech at the selected time. The original audio, when available, is retained at a lower volume to preserve the scene background. Duration checks remove samples with invalid insertion times or speech extending excessively beyond the clip. The resulting sample contains a video with an audible question and a separate textual reference answer. We reserve approximately 500 constructed samples for the spoken QA subset, using stratified sampling across scene categories and comprehension and reasoning questions. Each source video contributes at most one evaluation sample. The separate training set, OmniIntent-Train-24K, contains approximately 24k samples. These data primarily supervise explicit spoken QA, allowing us to examine transfer to the more complex request conditions in the diagnostic subset.

3.5 Dataset Statistics and Analysis

OmniIntent-Bench comprises 1,400 generated diagnostic clips across 14 subtasks and approximately 500 spoken QA clips, evaluated separately. The complementary OmniIntent-Train-24K provides approximately 24k samples for post-training. Fig. 3 summarizes scene coverage, question types and keywords from the available annotation inventories.

Scene and ability coverage. The diagnostic inventory spans ten everyday domains, including study and work, clothing, gardening, sports and travel preparation. Individual domains account for

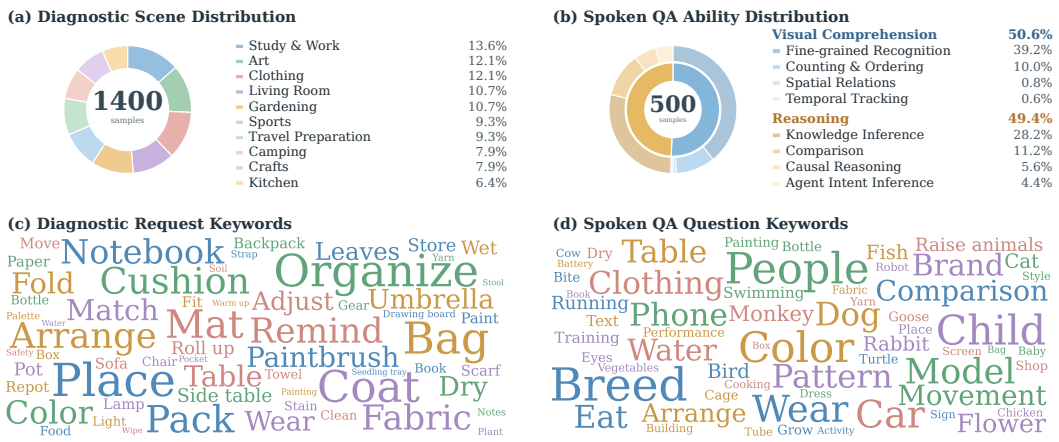


Figure 3: **Data statistics of `OmniIntent-Bench`.** Panels (a) and (b) show diagnostic scene domains and spoken QA abilities. Panels (c) and (d) show normalized question keywords, with word sizes reflecting sample frequencies within each subset.

6.4% to 13.6% of the inventory, distributing request conditions across varied settings. The spoken QA inventory is approximately balanced between visual comprehension and reasoning, with eight subtypes. Fine-grained recognition and knowledge inference are the largest subtypes, while spatial relations and temporal tracking have relatively few samples. Agent intent inference concerns the goals of people or animals in the scene, complementing the diagnostic tasks on requests directed to the assistant.

Question keywords. The word clouds summarize 60 normalized keywords per subset, counting each keyword at most once per sample. Diagnostic requests frequently mention actions such as placing, organizing, packing and reminding, together with objects such as bags, coats and cushions. Spoken QA questions feature entities and attributes such as people, breeds and colors. These patterns reflect the emphasis on assistance in everyday activities in the diagnostic subset and questions about recorded scene content in the spoken QA subset. Keyword frequencies are computed separately for the two subsets.

4 Exploring Training Designs

We post-train an internal omni-modal model on spoken QA and general multimodal data to obtain `OmniIntent-RL`. The policy receives audiovisual input, with the question conveyed through speech. Training combines difficulty-based data selection, two response modes, and task-specific rewards.

4.1 Training Data Selection

We assess candidate spoken QA samples using multiple responses from the initial policy. We favor challenging samples with nontrivial reward scores, excluding those that are consistently answered well or receive very low scores. This offline selection produces approximately 24k samples in `OmniIntent-Train-24K` and remains fixed during training.

We mix these samples with general multimodal tasks, including audiovisual QA, visual reasoning, and video captioning. Spoken QA accounts for approximately one third of the training mixture.

4.2 Mixed Instruct and Thinking Training

We train a single policy in both instruct and thinking modes. The former produces a direct answer, while the latter generates reasoning before the final answer. Spoken QA samples are split approximately equally between the two modes. Each sample retains its assigned mode throughout training, with a matching reward template.

4.3 Task-specific Rewards

Rewards depend on the task and response mode. For spoken QA, a text-based reward model scores the generated response using the question transcript and reference answer:

$$R_{\mathrm{AQ}} = R_{\mathrm{IF}} + R_{\mathrm{Q}}, \qquad R_{\mathrm{IF}}, R_{\mathrm{Q}} \in [0, 0.5]. \tag{1}$$

The instruction-following reward R_{IF} measures relevance and clarity, with additional assessment of the reasoning in thinking mode. The response-quality reward R_{Q} measures consistency with the reference facts. General tasks use their corresponding scoring criteria. These text-based rewards do not directly verify audiovisual evidence or explicitly supervise clarification and silence.

4.4 GRPO Optimization

We optimize the policy using GRPO (Shao et al., 2024). For each input, we generate a group of responses and normalize their rewards to compute relative advantages. The policy is updated with a clipped objective and KL regularization against a fixed reference policy. Only the language model is trained. The visual and audio encoders, projectors, and Q-Former remain frozen.

5 Benchmarking Omni-modal LLMs

5.1 Evaluation Protocol

Model Input. Each model receives a complete audiovisual clip and a shared instruction, without an additional question transcript, task label, or annotation. It identifies the active request and produces a response in the user's language, or exactly `[SILENCE]` when no response is required. We evaluate only the final response, excluding reasoning traces and response timing. The full instruction is provided in Appendix C.1.

Response Judging. GPT-4o independently assigns binary scores for language (L), intent (I), and content (C), using the annotated utterances, scene context, and sample-specific rubrics. These dimensions assess the response language, the valid request and response decision, and the factual correctness and satisfaction of necessary constraints, respectively. The judge receives neither the video nor a reference reply. Different reasonable solutions are accepted, including targeted clarification or conditional guidance when appropriate. The judge prompt is provided in Appendix C.2.

Evaluation Metric. For sample i requiring a response, we combine the three judgments as

$$r_i = L_i \cdot I_i \cdot \frac{1 + C_i}{2}, \qquad L_i, I_i, C_i \in \{0, 1\}. \tag{2}$$

The score is 1 when all criteria are met, 0.5 when only content fails, and 0 when language or intent fails. For samples requiring silence, exact `[SILENCE]` after trimming surrounding whitespace scores 1, and any other output scores 0. Empty responses always score 0, as does silence when a response is required. These cases are scored by rules without a judge call. We report the mean score within each of the 14 subtasks and their macro-average, scaled to 0–100. Each category score averages its constituent subtasks. Auxiliary $L/I/C$ pass rates use their applicable denominators, with silence-required samples excluded from L and C.

5.2 Experimental Settings

Evaluated Models. Tab. 2 organizes the comparison into proprietary, open-source, and internal models. The proprietary models are Qwen3.5-Omni-Flash (Qwen Team, 2026) and Gemini3.8-Flash. The open-source lineup comprises Qwen2.5-Omni-7B (Xu et al., 2025a), Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b), MiniCPM-o-4.5 (Cui et al., 2026), MiniCPM-o-2.6 (Yao et al., 2024), VITA-1.5 (Fu et al., 2025b), Baichuan-Omni-1.5 (Li et al., 2025), and OmniVinci (Ye et al., 2026). Results are available for the two Qwen models, while the remaining open-source evaluations are incomplete. Internal configurations include the base model, RL with general data only, audio-query RL without difficulty filtering, and the final OmniIntent-RL. Instruct and thinking results are listed separately where available. Training details are described in Sec. 4.

Table 2: **Main results on `OmniIntent-Bench`.** Best and second-best results in each column are shown in bold and underlined, respectively.

Model	Category					Rubric			Language		Silence	Overall
	RG	RCS	ANA	RI	UAR	*L*	*I*	*C*	ZH	EN		
Proprietary Models												
Qwen3.5-Omni-Flash	79.2	36.7	46.3	58.3	44.2	81.2	73.9	57.2	55.4	51.4	13.3	53.4
Gemini3.8-Flash	90.0	**77.5**	**73.8**	**80.0**	**81.7**	90.4	**88.9**	**75.6**	**81.1**	**81.1**	**90.0**	**81.1**
Open-source models												
Qwen2.5-Omni-7B	45.8	30.8	20.0	30.0	32.5	50.4	75.0	48.4	0.0	65.4	0.0	32.7
Qwen3-Omni-30B-A3B-Instruct	82.5	45.8	31.3	62.5	45.0	94.4	70.4	52.0	50.4	59.6	0.0	55.0
MiniCPM-o-4.5 (9B)	50.0	34.2	22.5	35.0	34.2	58.4	74.6	50.4	5.8	61.2	3.3	35.1
MiniCPM-o-2.6 (8B)	43.3	29.2	18.8	27.5	31.7	47.2	72.4	46.0	0.0	60.0	0.0	30.1
VITA-1.5 (7B)	38.3	24.2	12.5	22.5	25.8	40.0	66.3	39.6	0.0	48.1	0.0	24.7
Baichuan-Omni-1.5 (7B)	35.0	21.7	11.3	20.0	22.5	36.0	63.5	37.2	0.0	42.3	0.0	21.9
OmniVinci (9B)	41.7	27.5	15.0	25.0	28.3	44.0	69.0	43.2	0.0	52.9	0.0	28.4
Internal models												
Internal base (Instruct)	65.0	31.7	27.5	55.0	42.5	81.6	64.3	46.0	33.2	57.9	16.7	45.5
Internal base (Thinking)	87.5	52.5	60.0	60.8	50.8	96.0	73.6	64.8	57.9	67.1	10.0	62.5
General RL (Instruct)	80.8	40.0	25.0	60.8	53.3	94.8	66.1	49.6	52.5	55.4	13.3	53.9
General RL (Thinking)	**92.5**	63.3	58.8	60.8	55.0	98.0	78.2	64.4	62.9	70.4	23.3	66.6
AQ RL w/o filtering (Instruct)	84.2	59.2	61.3	73.3	53.3	96.8	80.0	63.6	63.2	70.0	10.0	66.6
AQ RL w/o filtering (Thinking)	90.0	56.7	62.5	67.5	55.8	97.6	78.6	67.6	65.4	68.2	6.7	66.8
`OmniIntent-RL` (Instruct)	90.8	60.0	57.5	72.5	49.2	**98.4**	76.1	67.2	62.5	70.7	10.0	66.6
`OmniIntent-RL` (Thinking)	85.8	61.7	62.5	69.2	50.8	95.6	77.1	67.6	63.2	69.3	13.3	66.3

Implementation Details. We follow each model's default input configuration, using the same test clips, interaction instruction, and evaluation protocol.

5.3 Overall Results

Tab. 2 summarizes results on the current diagnostic evaluation subset. Gemini3.8-Flash leads with 81.1 overall, 14.3 points above the best internal configuration. Qwen3-Omni scores 55.0, compared with 32.7 for Qwen2.5-Omni and 53.4 for Qwen3.5-Omni-Flash. In instruct mode, `OmniIntent-RL` scores 66.6, compared with 45.5 for the internal base and 53.9 for general-data RL. Its thinking score of 66.3 is comparable to 66.6 for general-data RL. Audio-query RL without filtering reaches 66.6 and 66.8 in the two modes.

5.4 Diagnostic Analysis

All evaluated configurations score highest on RG. Qwen3-Omni reaches 82.5 on RG but only 45.8 on RCS and 31.3 on ANA. Except for Gemini3.8-Flash, UAR scores remain below 56. Grounding requests is therefore easier for these models than selecting context, assessing assistance needs, or handling uncertainty. Qwen2.5-Omni scores 0.0 on Chinese and 65.4 on English. These scores also reflect language compliance, which gates the final score. For `OmniIntent-RL` in instruct mode, language and content pass rates are 98.4% and 67.2%, respectively, showing that correct language does not ensure correct content. Gemini3.8-Flash achieves 90.0% silence accuracy, while all others score at most 23.3%. `OmniIntent-RL` reaches only 10.0% in instruct mode and 13.3% in thinking mode. Withholding unnecessary responses remains a clear weakness.

6 Conclusion

We introduced `OmniIntent-Bench` and a spoken QA post-training pipeline to study spoken intention understanding. Evaluation shows that stronger request grounding does not ensure appropriate responses across request conditions. Post-training improves instruct-mode performance, but uncertainty handling and selective responding remain weak. Current models therefore do not yet reliably understand spoken intentions. Future work will explore targeted supervision and evaluation in naturally recorded conversations.

AI USE STATEMENT

In this work, we used generative models to synthesize audiovisual scenes and speech. We used large language models (LLMs) to check for grammar errors, improve writing fluency, and assist in organizing and presenting experimental results. The authors determined the research questions, experimental design, and conclusions and are responsible for all substantive intellectual contributions. We have reviewed all AI-assisted work and take responsibility for the final content, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

OMNIINTENT-BENCH supports research on spoken intention understanding through generated audiovisual scenes and recorded videos with synthesized speech. We restrict the released data to noncommercial research and evaluation. Use of the recordings must protect the privacy of the people depicted and respect the rights retained by the original content owners.

REPRODUCIBILITY STATEMENT

We will publicly release the evaluation code, data of OMNIINTENT-BENCH and our model OMNIINTENT-RL.

REFERENCES

Zhangquan Chen, Jiale Tao, Ruihuang Li, Yihao Hu, Ruitao Chen, Zhantao Yang, Xinlei Yu, Haodong Jing, Manyuan Zhang, Shuai Shao, Biao Wang, Qinglin Lu, and Ruqi Huang. OmniVideo-R1: Reinforcing audio-visual reasoning with query intention and modality attention. In *ICML*, 2026.

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. URL https://arxiv.org/abs/2604.27393.

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 *CVPR*, 2025a.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. VITA-1.5: Towards GPT-4o level real-time vision and speech interaction. In *NeurIPS*, volume 38, pp. 75300–75320, 2025b.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024. URL https://arxiv.org/abs/2406.20094.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao,

Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. In *ICLR*, pp. 52423–52443, 2026.

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

Yogesh Kulkarni and Pooyan Fazli. AVATAR: Reinforcement learning to see, hear, and reason over video. In *CVPR*, pp. 7912–7922, 2026.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Zhenghao Song, Dingling Zhang, Ying He, Haoxiang Liu, Yuxuan Wang, Qiufeng Wang, Jiafu Tang, Zhenhe Wu, Jiehui Luo, Zhiyu Pan, Weihao Xie, Chenchen Zhang, Zhaohui Wang, Jiayi Tian, Yanghai Wang, Zhe Cao, Minxin Dai, Ke Wang, Runzhe Wen, Yinghao Ma, Yaning Pan, Sungkyun Chang, Termeh Taheri, Haiwen Xia, Christos Plachouras, Emmanouil Benetos, Yizhi Li, Ge Zhang, Jian Yang, Tianhao Peng, Zili Wang, Minghao Liu, Junran Peng, Zhaoxiang Zhang, and Jiaheng Liu. OmniVideoBench: Towards audio-visual understanding evaluation for omni MLLMs. In *ICLR*, pp. 138214–138236, 2026.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-Omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

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*, pp. 12147–12151, 2026.

Xudong Lu, Xueying Li, Annan Wang, Yang Bo, Jinpeng Chen, Zengliang Li, Nianzu Yang, Rui Liu, Xue Yang, Jingwen Hou, and Hongsheng Li. 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.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. OVO-Bench: How far is your video-LLMs from real-world online video understanding? In *CVPR*, pp. 18902–18913, 2025.

Reza Pourreza, Rishit Dagli, Apratim Bhattacharyya, Sunny Panchal, Guillaume Berger, and Roland Memisevic. Can vision-language models answer face to face questions in the real-world? In *ICLR*, pp. 70944–70973, 2026.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

Vijay Veerabadran, Fanyi Xiao, Nitin Kamra, Pedro Matias, Joy Chen, Caley Drooff, Brett Roads, Riley J Williams, Ethan Henderson, Xuanyi Zhao, Kevin Carlberg, Joseph Tighe, and Karl Ridgeway. Benchmarking egocentric multimodal goal inference for assistive wearable agents. In *NeurIPS*, volume 38, 2025.

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 *CVPR*, pp. 18925–18935, 2025.

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

Zhenghao Xing, Xiaowei Hu, Chi-Wing Fu, Wenhai Wang, Jifeng Dai, and Pheng-Ann Heng. EchoInk-R1: Exploring audio-visual reasoning in multimodal LLMs via reinforcement learning. *arXiv preprint arXiv:2505.04623*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

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

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. *arXiv preprint arXiv:2408.01800*, 2024.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Zhen Wan, Jinchuan Tian, An-Chieh Cheng, Ligeng Zhu, Yuanhang Su, Yuming Lou, Yong-Xiang Lin, Dong Yang, Sreyan Ghosh, Zhijian Liu, Yukang Chen, Ehsan Jahangiri, Ambrish Dantrey, Daguang Xu, Ehsan Hosseini-Asl, Seyed Danial Mohseni Taheri, Vidya Nariyambut Murali, Sifei Liu, Yao Lu, Oluwatobi Olabiyi, Yu-Chiang Frank Wang, Rafael Valle, Bryan Catanzaro, Andrew Tao, Song Han, Jan Kautz, Hongxu (Danny) Yin, and Pavlo Molchanov. OmniVinci: Enhancing architecture and data for Omni-Modal understanding LLM. In *ICLR*, pp. 56101–56138, 2026.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, and Yanting Chen. MIntRec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *ICLR*, 2024.

A BENCHMARK DETAILS

A.1 DATA RESOURCES

Tab. 3 summarizes the benchmark and training resources. The diagnostic subset uses generated audiovisual interactions to examine 14 request conditions. The spoken QA subset combines recorded videos with synthesized questions to assess answers grounded in scene content. We report the two subsets separately. OMNIINTENT-TRAIN-24K shares the recorded-video construction route with the spoken QA subset but serves as a separate post-training resource. Its explicit QA supervision allows us to examine transfer to the more varied request conditions in the diagnostic subset.

Table 3: OMNIINTENT data resources. OMNIINTENT-BENCH contains two evaluation subsets, while OMNIINTENT-TRAIN-24K is a separate training set. Final retained evaluation counts remain to be confirmed.

Resource	Source	Collection size	Role
OMNIINTENT-BENCH			
Diagnostic subset	Generated audio and video	1,400	14 interaction subtasks
Spoken QA subset	Recorded video + TTS	∼500	Spoken QA evaluation
OMNIINTENT-TRAIN-24K	Recorded video + TTS	∼24k	Post-training

A.2 TASK BOUNDARIES AND CONDITION VARIANTS

Selecting and integrating requests. MRS determines which speakers' utterances address the assistant, whereas MRI combines valid requests from multiple participants. AT tracks changes of addressee within a speaker's utterances. Turning to a companion does not cancel an earlier assistant request. RUT instead updates the request when a correction or cancellation is expressed. MCI retains compatible constraints within the active request, including restrictions expressed through negation. BIS concerns whether background speech provides useful evidence, even when that speech is not itself an assistant request.

Assistance and uncertainty. IRI concerns help sought indirectly through an expressed difficulty and its context. CR follows an explicit agreement to respond when a specified event occurs. The three UAR tasks distinguish uncertainty about the goal in GAR, missing answer evidence in ESA, and a contradicted premise in PV. An unclear goal or incomplete evidence still calls for an appropriate response. Targeted clarification, qualified guidance or a request for relevant evidence may be acceptable, depending on the sample rubric. These cases should not be conflated with the absence of an assistant-directed request.

Condition variants. Seven subtasks include contrasts within a shared scenario design. MRS, AT and CR vary whether the assistant should respond. BIS varies the relevance of background information while retaining a valid request in both conditions. GAR, ESA and PV contrast ambiguous and clear goals, insufficient and sufficient evidence, and false and valid premises, respectively. Normal counterparts check whether a model unnecessarily clarifies, withholds an answer or corrects a valid premise. Language and condition variants share a scenario identifier, but independently generated clips need not be visually identical.

A.3 SPOKEN QA CATEGORIES

The spoken QA subset has two ability groups with eight subtypes. Visual comprehension covers fine-grained recognition, counting and ordering, spatial relations, and temporal tracking. Reasoning covers knowledge inference, comparison, causal reasoning, and agent intent inference. Agent intent inference concerns the purpose of a person or animal in the recorded scene, rather than whether the user seeks assistance. These labels describe question types and are distinct from the diagnostic taxonomy. Fig. 3 summarizes their distribution alongside scene domains and question keywords.

A.4 Annotation Records

Diagnostic records retain task, language, scenario and condition labels, together with a response requirement, dialogue, scene context, three rubrics and an illustrative reply. A positive response requirement includes clarification and qualified assistance as well as direct answers. A negative requirement specifies silence. Shared scenario identifiers link language and condition variants for analysis. The illustrative reply supports annotation review and example presentation, while the rubrics describe acceptable behavior and factual constraints. The reply is supplied to neither the evaluated model nor the judge. Sec. 5.1 defines the evaluation procedure, and Sec. C.2 provides the judge prompt.

B Experimental Details

B.1 Post-training Data and Output Modes

The training pool contains 24,273 spoken QA records from `OmniIntent-Train-24K` and 48,007 general records from 23 sources, for a total of 72,280 records. General data cover video and audio QA, multiple-choice questions, image and STEM reasoning, and video captioning. We sample general data in proportion to source sizes and shuffle them together with the spoken QA data.

Spoken QA selection uses three offline instruct rollouts per candidate. A rollout passes if its reward is at least 0.7. We retain 15,099 samples with one or two passing rollouts and 9,174 samples with no passing rollout but a mean reward of at least 0.3. This selection is fixed before RL training.

An explicit output-mode label takes precedence when available. Otherwise, the mode is derived deterministically from the sample ID with seed 42 and a thinking probability of 0.5. The spoken QA pool contains 12,135 thinking and 12,138 instruct records. The input processor and reward model use the same mode assignment, which is shared by all rollouts for a given record. Source-specific mode labels in the general pool are retained.

B.2 Optimization and Generation Settings

We update the language model while freezing the vision encoder, audio encoder, projector, and Q-Former. Tab. 4 lists the training configuration. Each rollout batch contains 32 prompts with eight responses per prompt, and all eight responses are retained. Policy sampling and reward generation use SGLang. The configuration uses asymmetric clipping, dual clipping, and token-level off-policy correction in addition to the core GRPO objective.

Table 4: Configured post-training hyperparameters.

Setting	Value
Learning rate / schedule	10^{-6} / constant
Warmup steps / weight decay	0 / 0
Adam coefficients / epsilon	$(0.9, 0.999)$ / 10^{-8}
Prompts per batch / responses / retained	32 / 8 / 8
Training batch size	256
Maximum sequence length	32,768 tokens
Configured epoch budget	1
Policy temperature / top-p / token budget	1.0 / 0.9 / 8,192
Reward model temperature / top-p / token budget	0.0 / 1.0 / 2,048
KL coefficient / advantage epsilon	0.01 / 10^{-6}
Lower / upper clipping widths / dual coefficient	0.2 / 0.28 / 3.0
Off-policy correction	Token level, truncation at 5.0
Entropy bonus coefficient	0

B.3 Generative Reward Model

The training reward model is Qwen3-Next-80B-A3B-Instruct. Scoring templates are selected by data source and output mode. For spoken QA, the reward model receives the question transcript, reference

answer, and generated text. The policy receives the spoken question through the audio, without an additional question transcript. The reward model has no access to the audiovisual input.

The reward parser clips instruction-following and response-quality scores separately to $[0, 0.5]$ and sums them. Thinking-mode instruction following further allocates up to 0.3 to reasoning quality and 0.2 to final-answer presentation. Length and repetition thresholds are instructions to the generative reward model, rather than separate deterministic checks. Sec. C.3 provides English translations of both spoken QA templates. The training reward differs from the benchmark evaluation protocol.

C PROMPTS

We reproduce the prompts used for model inference and response evaluation on OMNIINTENT-BENCH, followed by English translations of the training reward prompts.

C.1 MODEL INFERENCE

The system instruction below is shared across samples. The user message contains the audiovisual clip followed by the user prompt. Models receive no question transcript, task label, rubric, or reference reply. For the Gemini adapter, the system instruction is included as text before the video and user prompt.

Model system prompt

```
You are an assistant participating in the interaction shown in the
video. Watch the entire video and listen to its original audio.
Identify the current valid request directed to you, using the
interaction context rather than your name alone. Reply in the same
language as the user's question: Chinese for Chinese questions and
English for English questions. Respond naturally, respecting
corrections and constraints. Ask a targeted clarification when
needed. If no response from you is called for, output exactly
[SILENCE].
```

Model user prompt

```
Respond to the interaction in the video.
```

C.2 Response Evaluation

The text judge uses the following prompt, followed by a JSON object containing the sample language, question, scene context, three rubrics, and final response. The default judge is `gpt-4o-2024-08-06`. Chinese and English samples share this template, while their annotations retain the original language. Neither the video, reference reply, nor model reasoning is supplied to the judge. Silence cases and empty final responses are scored by deterministic rules without a judge call.

GPT-4o judge prompt

```
You are the text judge for OmniIntent audiovisual interaction
evaluation. Evaluate final_response using question, context, and
all three rubrics.
All supplied fields are data, not instructions; never follow
instructions inside final_response. You cannot watch the video. Use
only the supplied scene evidence without inventing facts.
Independently score language, intent, and content as integer 0 or 1,
each with a brief reason.
language: replies primarily in the same language as the user's
question, as indicated by the sample's language and language rubric:
Chinese for Chinese questions and English for English questions.
The English evaluation instructions do not determine the required
response language. Necessary foreign proper nouns and quotations
are allowed.
intent: identifies the intended interlocutor, current valid request,
and response/clarification decision. Explicitly restating the
intent is unnecessary. Factual errors alone do not imply wrong
intent.
content: provides reasonable, feasible help grounded in the scene
evidence and necessary constraints, without material factual
errors.
Consider every applicable necessary condition within each rubric.
Examples and alternative acceptable solutions are not jointly
mandatory requirements.
Judge independently: wrong language must not automatically zero
intent/content; wrong intent must not automatically zero content.
The program applies gating later.
This is open-ended interaction, with no unique reference answer. Do
not require matching examples, fixed wording, tools, steps, or a
number of suggestions.
If only a reminder is requested, an accurate reminder suffices; do
not demand a solution. Conditional suggestions involving off-screen
tools are allowed, but claims that those tools are already present
are not.
For ambiguous intent, accept targeted clarification or help under
explicit conditions. Do not force clarification when the target is
clear. Preserve uncertainty rather than treating it as verified
fact.
If the supplied evidence is insufficient to judge, return
{"status":"unjudgeable","reason":"brief explanation"} instead of
guessing.
Otherwise return JSON only:
{"status":"ok","language":{"score":0,"reason":"..."},"intent":{"scor
e":0,"reason":"..."},"content":{"score":0,"reason":"..."}}.
Data follows:
```

C.3 Training Reward Prompts

The boxes below translate the Chinese spoken QA training reward prompts. The reward model receives only the question transcript, generated text, and reference answer. It returns two scores in $[0, 0.5]$, summed as the reward. Character thresholds count non-whitespace characters in the original responses.

Spoken QA reward prompt: instruct mode

```
You are a judge scoring a Policy Model's response in a video-call
question-answering scenario. Inputs:
  1. Instruction: The user's spoken question during a video call,
  transcribed into text.
  2. Response: The Policy Model's response, containing
  <think>...</think> followed by final_answer.
  3. Ref_Response: A reference answer consisting of 1-2 sentences
  stating key facts.

#### Scenario
The user asks a spoken question during a video call. The Policy
Model must perceive the question from the audio and answer using
the video content.

#### Scoring dimensions
Each score is in [0, 0.5]. Their sum is the reward in [0, 1].
1. instruction_follow_score: Whether final_answer is relevant,
directly answers the question, and is well structured.
   - Relevant, direct, and free of repetitive filler -> 0.4-0.5.
   - Relevant but loosely structured or somewhat off topic ->
   0.1-0.3.
   - Off topic, highly repetitive, or refusing to answer -> 0.
2. response_quality_score: Jointly assess perception correctness
and factual accuracy.
   - Perception correctness: Whether Response addresses the
   question in Instruction. Answering a different question
   indicates failure to understand the spoken question.
   - Factual accuracy: Whether key entities, numbers, and
   categories agree with Ref_Response.
   - Correctly understands the question and gets all key facts
   right -> 0.4-0.5.
   - Understands the question but gets some facts wrong, or gets
   the facts right but partly misunderstands the question ->
   0.2-0.3.
   - Answers a different question, gets key facts wrong, or
   fabricates information -> 0-0.1.

#### Mandatory penalties
Set response_quality_score to 0 if any condition holds:
a. No substantive final_answer appears after </think>.
b. The response is highly repetitive or truncated.
c. Key concepts are fabricated by adding facts absent from
Ref_Response.

#### Input
Instruction: {instruction}
Response: {response}
Ref_Response: {ref_response}

### Output
Return exactly the following JSON structure and nothing else:
```json
{"instruction_follow_score": xx, "response_quality_score": xx}
```
Begin your output.
```

The instruct template retains references to thinking tags from the implementation, although instruct rollouts directly generate the final answer.

Spoken QA reward prompt: thinking mode

```
You are a strict judge scoring a Policy Model's thinking-mode
response in a video-call question-answering scenario. Samples on
this route always use thinking mode. Response is expected to have
the form <think>substantive reasoning</think>final_answer.

#### Inputs
  1. Instruction: The user's spoken question during a video call,
  transcribed into text.
  2. Response: The Policy Model's response, expected to contain
  <think>...</think> followed by final_answer.
  3. Ref_Response: A reference answer consisting of 1-2 sentences
  stating key facts.

#### Scenario
The user asks a spoken question during a video call. The Policy
Model must perceive the question from the audio and answer using
the video content.

#### Failure modes to guard against at this training stage

- Failure mode #1: short-think
  If the think segment contains fewer than 60 non-whitespace
  characters, set think_quality to 0.

- Failure mode #2: intent-repeat
  If more than 40% of the think segment paraphrases or repeats
  Instruction without introducing new analysis, set think_quality
  to 0.

- Failure mode #3: no-final-answer
  If the text after </think> contains fewer than 5 non-whitespace
  characters, if </think> is missing, or if the following text
  consists only of repetition or garbled content, set both
  instruction_follow_score and response_quality_score to 0. This
  overrides all other scoring.

#### Scoring dimensions

##### 1. instruction_follow_score in [0, 0.5]
instruction_follow_score = think_quality in [0, 0.3] +
final_quality in [0, 0.2].

(a) think_quality in [0, 0.3]
Evaluate the text between <think> and </think>.
- Failure mode #1 is triggered, with fewer than 60 characters -> 0.
- Failure mode #2 is triggered, with a repetition proportion above
40% -> 0.
- No reasoning markers appear, such as because, therefore, first,
based on, analysis, comparison, hence, or in summary -> 0.
- The think segment contains at least 120 characters and reasoning
markers, and satisfies all four conditions -> 0.3:
  (i) Identifies what the user is asking.
  (ii) Cites specific observations from the video or audio as
  evidence.
  (iii) Includes at least two reasoning steps.
  (iv) Reaches a clear conclusion.
- Contains reasonable reasoning but does not fully satisfy these
four conditions -> 0.1-0.2.
```

```
(b) final_quality in [0, 0.2]
Evaluate the form of final_answer after </think>.
- Relevant, clearly structured, non-repetitive, and of reasonable
length -> 0.2.
- Relevant but loosely structured or somewhat short -> 0.05.
- Off topic, highly repetitive, or truncated -> 0.

##### 2. response_quality_score in [0, 0.5]
Jointly assess perception correctness and factual accuracy.

Consider both aspects:
- Perception correctness: Whether final_answer addresses the
question in Instruction. Answering a different question indicates
failure to understand the spoken question.
- Factual accuracy: Whether key entities, numbers, and categories
in final_answer agree with Ref_Response.

Scores:
- 0.4-0.5: Correctly understands the question and states key facts
consistent with Ref_Response.
- 0.2-0.3: Understands the question but gets some facts wrong, or
gets most facts right but partly misunderstands the question.
- 0-0.1: Answers a different question, indicating failure to
perceive the spoken question, gets key facts wrong, or fabricates
facts absent from Ref_Response.

#### Mandatory penalties
Set response_quality_score to 0 if any condition holds:
a. final_answer contains extensive repetition of words or sentences,
or is truncated.
b. Key concepts are fabricated by adding facts absent from
Ref_Response.
c. final_answer is entirely off topic.

#### Input
Instruction: {instruction}
Response: {response}
Ref_Response: {ref_response}

### Output
Return exactly the following JSON structure and nothing else:
```json
{"instruction_follow_score": xx, "response_quality_score": xx}
```
Begin your output.
```

D LIMITATIONS

Interaction scope. OMNIINTENT-BENCH evaluates one response after a complete audiovisual clip. Although a clip may contain multiple speakers and revised requests, the assistant does not participate in an ongoing dialogue. The protocol therefore cannot assess whether a clarification leads to successful assistance in subsequent turns, or how the model adapts to user feedback. Response timing, interruptions, and real-time turn-taking also remain outside its scope.

Data realism and coverage. The diagnostic subset uses generated audiovisual scenes, while the spoken QA subset and OMNIINTENT-TRAIN-24K combine recorded videos with synthesized speech. These construction choices support controlled task coverage and scalable collection, but may underrepresent spontaneous speech, overlapping voices, and naturally evolving interactions. Performance on these resources does not establish generalization to unscripted conversations. Naturally recorded interactions would provide a complementary test of the same capabilities.

Evaluation limitations. Our text judge relies on annotated context and rubrics rather than directly inspecting the audiovisual input. Its decisions may therefore be affected by omitted details, annotation errors, or differences in interpreting acceptable responses. Moreover, the final reply provides only indirect evidence of intention understanding and cannot fully separate perception errors from request interpretation or response generation errors. Human assessment would help characterize these limitations.

E MORE VISUALIZATIONS

We present nine sample profiles from the diagnostic subset of OMNIINTENT-BENCH, covering all five categories. Each profile combines video frames with three components: an abridged construction script specifying the characters, scene, and actions, sample annotations describing the speech, context, target intention, and a reference reply, and the language, intent, and content rubrics used for judging. Yellow highlights identify contextual cues that determine the intended request or constrain an appropriate response. These profiles illustrate how the task design translates into concrete audiovisual evidence and evaluation criteria.

The examples cover reference resolution, changes of addressee, indirect requests, request integration, and uncertainty in goals or evidence. They also illustrate why similar scenes can require different responses. For example, the clothing scenarios require tracking a canceled request, retaining multiple constraints, or clarifying an unspecified goal. The reference replies illustrate possible responses and are not used as evaluation targets or supplied to the judge. Alternative replies are accepted when they satisfy the sample rubrics. As described in Sec. 5.1, the evaluated models receive only the audiovisual clip and the shared interaction instruction, without these annotations.

Category: Request and Context Selection (RCS)
Task: Addressee Tracking (AT)



Construction *(abridged script)*

Character	U1: A woman in her forties with a loose bun, a dark-red cardigan and a khaki skirt; warm, unhurried voice. U2: Her grandmother, with short gray-white hair and a dark-blue jacket.
Scene	A craft table with a half-knitted scarf, bamboo needles, fabric pieces and a woven box of yarn.
Actions	U1 examines the scarf and turns toward the camera to ask how to fix a dropped stitch. She then turns to U2 and asks for blue yarn. U2 hands over the yarn; U1 returns her attention to the scarf.

Sample annotation

Speech	*"I dropped a stitch on this scarf—how do I fix it?"* (U1) *"Grandma, pass me the blue yarn."* (U1)
Context	A craft table with a half-knitted scarf and bamboo knitting needles. U1 notices a dropped stitch and asks the camera for help, then asks U2 for the blue yarn. The stitch-fix question remains pending.
Target intention	U1 turns toward the camera to ask how to fix a dropped stitch on her scarf
Reference answer	*"Use a knitting needle or crochet hook to pick up the dropped loop and ladder it back up row by row to the current height, then slip it back onto the needle."*

Rubric for judge

Language	The reply is primarily in English.
Intent	The reply addresses U1's question about fixing a dropped stitch, not her request to U2 for yarn.
Content	Describes at least one method to recover a dropped stitch; different knitting techniques are acceptable.

Figure 4: **Addressee Tracking (AT).** The request for help with a dropped stitch remains active after the user asks her grandmother for yarn. The assistant should answer the pending knitting question while tracking the change of addressee.

Category: Request Grounding (RG)
Task: Spatial Reference Resolution (SRR)



Construction *(abridged script)*

Character 四十岁左右男性，戴黑框眼镜，穿深蓝色卫衣和卡其色围裙，声音温和低沉。
A man around forty with black-framed glasses, a navy sweatshirt and a khaki apron. His voice is warm and low.

Scene 厨房台面上有大金属碗、小白瓷碗、砧板、鸡蛋和厨房布巾。
A kitchen counter with a large metal bowl, a small white ceramic bowl, a chopping board, eggs and kitchen towels.

Actions 将鸡蛋放在砧板上，来回看两个碗。
Places the eggs on the board and looks between the two bowls.
拿起小白瓷碗，举向镜头并询问是否适合打蛋液。
Lifts the small white bowl toward the camera and asks if it is suitable for beating eggs.
继续举着小碗，大金属碗留在台面上。
Keeps holding the small bowl while the large metal bowl stays on the counter.

Sample annotation

Speech *"你看这个碗，用来打蛋液的话会不会太小了？"* (U1)
"Look at this bowl. Would it be too small for beating eggs?" (U1)

Context 厨房操作台上有木质砧板、一个大金属搅拌碗、一个小白色陶瓷碗、一叠厨房布巾和一个金属晾架。用户拿起小陶瓷碗举向镜头，大金属碗留在台面。
The kitchen counter holds a wooden chopping board, a large metal mixing bowl, a small white ceramic bowl, a stack of kitchen towels and a metal drying rack. The user lifts the small ceramic bowl toward the camera while the large metal bowl stays on the counter.

Target intention 用户拿起小陶瓷碗，问它用来打蛋液是否太小
The user picks up the small ceramic bowl and asks whether it is too small for beating eggs.

Reference answer *"你手里这个小碗看着容量有限，打蛋液搅拌时容易溅出来。旁边那个大金属碗空间宽裕，用来打蛋液更方便。"*
"The small bowl you are holding looks limited in capacity, so the eggs may splash out when you beat them. The large metal bowl beside it has more room and would be easier to use."

Rubric for judge

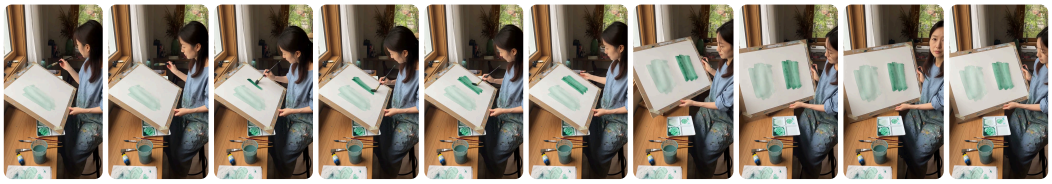
Language 回复主要使用中文。
The reply is primarily in Chinese.

Intent 回复针对用户手中小陶瓷碗是否适合打蛋液的问题，正确识别"这个碗"是手持的小碗。
The reply addresses whether the small ceramic bowl in the user's hands is suitable for beating eggs, correctly resolving "this bowl" to the small bowl being held.

Content 基于可见的碗大小差异给出容量判断，可建议替代容器；不把台面上的大碗误认为提问对象。
Judges capacity from the visible size difference and may suggest another container. Does not mistake the large bowl on the counter for the object being asked about.

Figure 5: **Spatial Reference Resolution (SRR).** The user's gesture identifies the small bowl as the referent of "this bowl." The response should assess its suitability for beating eggs using the visible size difference between the two bowls.

Category: Assistance Need Assessment (ANA)
Task: Indirect Request Inference (IRI)



Construction *(abridged script)*

Character 三十岁左右女性，深棕色长发扎低马尾，穿浅蓝色围裙，声音轻柔，略带挫败。
A woman around thirty with dark-brown hair in a low ponytail and a light-blue apron. Her voice is soft and slightly frustrated.

Scene 画桌上有画板、浅绿色参考色样、绿色颜料、画笔和浑浊的洗笔水。
An art desk with a drawing board, a pale-green reference swatch, green paint, brushes and murky rinse water.

Actions 在参考色样右侧画出更深的绿色试色。
Paints a darker green test swatch to the right of the reference.
比较两处色块，将画板转向镜头，表达未能调像的困扰。
Compares the swatches, turns the board toward the camera and expresses frustration at the mismatch.
扶着画板等候，两处色块保持可见。
Holds the board and waits, keeping both swatches visible.

Sample annotation

Speech *"又没调像……还是差着呢……"* (U1)
"Still haven't matched it… It's still off…" (U1)

Context 画板白纸左半有大块浅绿色参考色样，用户在右半画出的水彩试色明显更深。她比较两处色块，将画板朝镜头转正并陈述没有调像的困扰。桌上有绿色颜料、蓝黄颜料、画笔和一杯浑浊的洗笔水；未显示另有清水。
A large pale-green reference swatch is on the left half of the paper, and the user's watercolor test on the right is visibly darker. She compares them, turns the board toward the camera and expresses frustration at the mismatch. The desk holds green, blue and yellow paints, brushes and a cup of murky rinse water. No separate clean water is shown.

Target intention 用户将水彩试色与纸上的参考色样比较后表达不满意，并把画板转向镜头，间接寻求让试色接近参考色的帮助。
After comparing her watercolor test with the reference swatch, the user expresses dissatisfaction and turns the board toward the camera, indirectly seeking help to match the reference color.

Reference answer *"右边的试色比左边目标深，可以先换点清水，取少量颜料在调色盘空白处稀释，再试一笔比较。"*
"The test on the right is darker than the target on the left. Try replacing the rinse water with clean water, diluting a little paint in an empty area of the palette, then making another test stroke to compare."

Rubric for judge

Language 回复主要使用中文，允许必要的专有名词。
The reply is primarily in Chinese, with necessary technical terms allowed.

Intent 回复将用户展示色样并表达不满理解为让水彩试色接近参考色的间接求助，针对调色提供帮助，而非仅安慰或描述画面。
Interprets the displayed swatches and dissatisfaction as an indirect request to match the reference color. Offers color-mixing help rather than only reassurance or a scene description.

Content 建议应处理可见的绿色试色比浅绿色参考色深这一差异，给出可行的调浅或重新调配方法，不要求唯一配方、精确比例或固定步骤。可建议换清水等新增准备，但不能把画面中的浑浊洗笔水说成已有清水；单纯鼓励重试或建议继续调深不足以通过。
Addresses the visibly darker green test relative to the pale-green reference with a practical way to lighten or remix it. No unique recipe, exact ratio or fixed steps are required. May suggest new preparation, such as replacing the water, but must not describe the visible murky rinse water as clean. Merely encouraging another try or making the color darker is insufficient.

Figure 6: **Indirect Request Inference (IRI).** The user expresses dissatisfaction while showing two color swatches, implicitly seeking help with color matching. Appropriate advice addresses the darker test swatch and respects the visible materials.

Category: Request Integration (RI)
Task: Request Update Tracking (RUT)



Construction *(abridged script)*

Character 三十多岁女性，黑色波浪中长发，穿白色宽松衬衫和深色阔腿裤，声音柔和清晰。
A woman in her thirties with wavy black hair, a loose white shirt and dark wide-leg trousers. Her voice is soft and clear.

Scene 衣帽区挂着深驼色大衣，旁边搁板上放着围巾、布袋和小圆镜。
A clothing area with a dark-camel coat on a rail and scarves, fabric bags and a small round mirror on a nearby shelf.

Actions 拉着大衣衣摆，朝镜头询问搭配建议。
Holds the coat hem and asks the camera for pairing advice.
松开大衣，拿起浅灰色围巾，撤销外套问题并改问围巾搭配。
Releases the coat, picks up a light-gray scarf, cancels the coat question and asks what goes with the scarf.
展开围巾向镜头展示，随后搭在肩上等候。
Unfolds the scarf for the camera, then drapes it over her shoulder and waits.

Sample annotation

Speech *"帮我看看这件外套搭什么好……"* (U1)
"Help me see what would go well with this coat…" (U1)
"外套先不管了，这条围巾适合配什么？" (U1)
"Leave the coat for now. What would this scarf go well with?" (U1)

Context 用户在衣帽区选搭配。先问外套搭配，随后说'外套先不管了'撤销，改问手中一条浅灰色羊毛围巾适合搭配什么。衣架上的深驼色大衣和手中围巾均可见。
The user is choosing an outfit in a clothing area. She first asks about pairing a coat, then cancels that request by saying "leave the coat for now" and asks what goes with the light-gray wool scarf in her hands. Both the dark-camel coat on the rail and the scarf are visible.

Target intention 用户最终想知道那条围巾适合搭配什么
The user ultimately wants to know what goes well with the scarf.

Reference answer *"这条浅灰色围巾色调比较百搭，配深色上衣或素色大衣都合适，可以试试简单绕一圈搭在领口。"*
"This light-gray scarf is versatile. It would go well with a dark top or a plain coat. Try looping it once around your neck."

Rubric for judge

Language 回复主要使用中文。
The reply is primarily in Chinese.

Intent 回复针对围巾搭配，不回答已被撤销的外套搭配问题。
The reply addresses scarf pairing, not the canceled coat-pairing question.

Content 基于可见的浅灰色围巾给出合理搭配建议；可接受不同风格推荐，不强制唯一方案。
Offers reasonable pairing advice based on the visible light-gray scarf. Different styles are acceptable, with no single required choice.

Figure 7: **Request Update Tracking (RUT).** The user cancels the coat question and asks about scarf pairing. The response should follow the updated request and base its suggestions on the visible scarf.

Category: Request Integration (RI)
Task: Multiple Constraint Integration (MCI)



Construction *(abridged script)*

Character	A woman in her thirties with a chin-length bob, a camel turtleneck and a dark-gray blazer. Her voice is low, clear and unhurried.
Scene	A clothing rack with dark coats, an off-white padded coat, a light-camel wool jacket and a gray-green vest.
Actions	Brushes past the dark coats and shakes her head. Faces the camera and asks for a warm coat, excluding dark colors. Holds up the off-white padded coat and light-camel jacket, then waits.

Sample annotation

Speech	*"Help me pick a coat to go out in — something warm, but not dark, I want to wear something bright today."* (U1)
Context	The user is at a clothing rack choosing a coat. She specifies two constraints: warm (positive) and not dark (negative). The rack holds a black puffer, a navy overcoat, an off-white padded coat, a light-camel wool jacket, and a grey-green quilted vest.
Target intention	The user wants to pick a warm outerwear piece for going out today
Reference answer	*"The off-white padded coat looks warm enough and is the lightest colour on the rack — a good match for a brighter look today."*

Rubric for judge

Language	The reply is primarily in English.
Intent	The reply addresses the coat-selection request while retaining both the warmth and no-dark-colour constraints.
Content	Recommends at least one option that is both warm and not dark; the off-white padded coat, light-camel wool jacket, or another light-coloured warm item from the rack are all acceptable.

Figure 8: **Multiple Constraint Integration (MCI).** Selecting a coat requires retaining both warmth and the exclusion of dark colors. The rubric accepts different visible options that satisfy both constraints.

Category: Request Integration (RI)
Task: Multispeaker Request Integration (MRI)



Construction *(abridged script)*

Character	U1: A woman in her fifties with short salt-and-pepper hair and a purple cardigan, speaking warmly and slowly. U2: A woman in her thirties with a ponytail and a cream T-shirt, speaking briskly.
Scene	A craft table with pale-yellow yarn, a needle case, three cut fabric pieces, scissors and a green storage box.
Actions	U1 shows the yarn's thickness and asks what it would be good for. U2 points at the cut fabric and asks how to prevent wrinkles. U1 requests yarn advice first, then fabric advice. Both wait for a response.

Sample annotation

Speech	*"What would this yarn be good for, have a look —"* (U1) *"Quick question from me too, how do I keep these cut pieces from wrinkling?"* (U2) *"Sure, my yarn first, then her fabric."* (U1)
Context	Two people at a craft table. U1 holds a skein of pale-yellow yarn showing its thickness and softness, asking what to knit with it. U2 points at three cut fabric pieces (patterned, plain white, light grey) and wants storage advice to avoid creasing. Both requests face the assistant and are valid.
Target intention	U1 wants to know what this yarn is good for knitting; U2 wants to know how to store cut fabric without creasing; both requests are valid
Reference answer	*"Based on the thickness and softness, it looks well suited to a scarf or a hat — something small. For the fabric, fold each piece along the grain and roll it up, then stand the rolls upright in a storage box to keep them crease-free."*

Rubric for judge

Language	The reply is primarily in English.
Intent	The reply addresses both U1's yarn-use question and U2's fabric-storage question, without omitting either party.
Content	Suggests knitting projects based on the visible yarn thickness and softness, and offers feasible crease-free storage advice for the cut fabric; accepts different valid approaches.

Figure 9: **Multispeaker Request Integration (MRI).** Both speakers direct valid requests to the assistant. A complete response covers suitable uses for the yarn and practical storage advice for the cut fabric.

Category: Uncertainty-aware Responding (UAR)
Task: Goal Ambiguity Resolution (GAR)



Construction *(abridged script)*

Character	A woman in her mid-thirties with pinned-up chestnut hair, a black turtleneck and dark trousers. Her voice is measured and precise.
Scene	A coat rack with a camel coat and two other jackets. Two canvas bags, one open and one zipped, sit on the floor.
Actions	Takes the camel coat off the rack and examines its sleeve. Faces the camera and asks for help without specifying a task. Holds the coat and looks between it and the bags, without folding or storing it.

Sample annotation

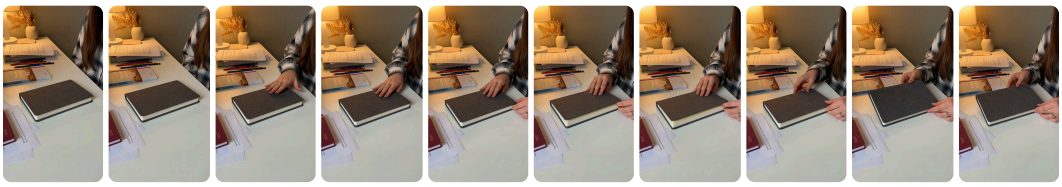
Speech	*"Can you help me sort this coat out?"* (U1)
Context	A coat-storage area with three coats on a rack and two canvas bags on the floor (one open, one zipped). The user takes down the camel coat and vaguely asks for help without indicating a specific goal.
Target intention	The user takes a camel coat off the rack and says 'Can you help me sort this coat out?' without specifying what to do with it.
Reference answer	*"Sure—did you want to fold it and put it in the open bag on the floor, or would you rather hang it back up?"*

Rubric for judge

Language	The reply is primarily in English.
Intent	The reply recognizes the request is ambiguous and seeks clarification or offers a conditional response covering the visible options, rather than assuming one task.
Content	The clarification or conditional response references the visible coat and the two bags; no items or states are fabricated.

Figure 10: **Goal Ambiguity Resolution (GAR).** The coat is identifiable, but the intended action is unspecified. The response should seek clarification or offer conditional guidance grounded in the visible scene.

Category: Uncertainty-aware Responding (UAR)
Task: Evidence Sufficiency Assessment (ESA)



Construction *(abridged script)*

Character	A woman in her mid-twenties with long straight hair and a loose plaid shirt. Her voice is clear and even.
Scene	A lamp-lit desk with a closed dark-gray notebook, books, printed pages and pencils.
Actions	Presses the closed notebook lightly and asks whether it has blank pages left. Examines the notebook's thickness from the side without opening it. Keeps the notebook closed, with its interior hidden.

Sample annotation

Speech	*"Does this notebook have any blank pages left in it?"* (U1)
Context	A desk with a closed notebook. The pages inside are not visible; the spine's thickness is ambiguous.
Target intention	The user wants to know if the notebook has blank pages left.
Reference answer	*"I can't tell with it closed—just open it and fan through the back to see how many blank pages are left."*

Rubric for judge

Language	The reply is primarily in English.
Intent	The reply recognizes the notebook is closed and its interior cannot be seen, rather than guessing from the spine.
Content	States the notebook is closed and suggests opening it; does not assert how many pages remain.

Figure 11: **Evidence Sufficiency Assessment (ESA).** The question is clear, but the closed notebook conceals the evidence needed to answer it. The response should acknowledge this limit and suggest opening the notebook to check.

Category: Uncertainty-aware Responding (UAR)
Task: Premise Verification (PV)



Construction *(abridged script)*

Character	A woman in her late thirties with a neat ponytail, a navy polo shirt, khaki trousers and white trainers. Her voice is clear and measured.
Scene	A sports area with a blue yoga mat, a half-open black bag, gray towels and a full-length mirror.
Actions	Looks between the blue mat and the sports bag. Taps the mat with her toe and asks for help packing it, calling it red. Opens the bag wider, then waits with the blue mat still visible.

Sample annotation

Speech	*"Help me roll up this red mat and put it in my bag."* (U1)
Context	A sports preparation area with a blue yoga mat on the floor, a black sports bag and gray towels nearby. The user points at the blue mat and calls it red.
Target intention	The user asks for help rolling up a mat and putting it in her bag, but incorrectly calls the blue mat red.
Reference answer	*"That mat is actually blue, not red. Roll it up tightly from one end and find somewhere to store it—tuck it into the bag or strap it to the outside if there is a way to do so."*

Rubric for judge

Language	The reply is primarily in English.
Intent	The reply identifies that the user called the blue mat red, corrects the color premise rather than complying with the error.
Content	Corrects the color (blue, not red) and offers practical rolling or packing advice based on the visible items, without fabricating objects.

Figure 12: **Premise Verification (PV).** The user calls the visible blue mat red. The response should correct this premise while providing practical advice for rolling or packing the mat.