OMNI DEMAND UNDERSTANDING: A BENCHMARK FOR CONTEXTUAL USER-INTENT INFERENCE IN MULTIMODAL INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural audio–visual interaction is emerging as an important interface for AI assistants, allowing users to communicate through speech and vision rather than carefully composed text prompts. However, existing benchmarks of interactive capabilities still focus primarily on response quality, leaving a more fundamental question underexplored: *can a model correctly infer the user's underlying demand from a complex multimodal interaction?* Real-world user demands are often underspecified in speech and must be inferred from multimodal cues and dialogue history. This inference is further complicated by ambiguous or disfluent expression and noisy acoustic environments. Conversely, request-like speech may not constitute a demand to the assistant, leading to false triggers. We establish **Omni Demand Understanding (ODU)** as a distinct multimodal contextual inference problem: given an interaction stream, a model must detect whether a user demand is present and infer the user's intent from multimodal and conversational context. ODU evaluates this capability along five dimensions, covering both single-turn and multi-turn interactions. We construct ODU-Bench using a challenge-driven taxonomy, taxonomy-guided agentic video generation, and human-recorded interactions, followed by media-grounded annotation and human verification. We evaluate 14 native audio and audio–visual MLLMs. Even Gemini 3.1 Pro, the strongest model by audio–visual aggregate score, recovers only 44.7% of the key contextual information needed to interpret user demands. Moreover, 11 of the 14 models exhibit false-trigger rates above 50% on non-demand scenarios. These results reveal a systematic capability gap in current MLLMs' ability to infer contextual user demands. We hope ODU can establish the evaluation of a previously underexplored yet essential capability in multimodal interaction: correctly understanding user demands before generating an appropriate response. See demos at `https://odubench.github.io`

1 INTRODUCTION

Multimodal large language models (MLLMs) that jointly perceive audio, vision, and language are increasingly serving as conversational assistants (Google DeepMind, 2026; Seed, 2026; Qwen Team, 2026c; Xu et al., 2025b;a; Cui et al., 2026; Hurst et al., 2024; AI et al., 2025; Deshmukh et al., 2026; Tang et al., 2025). Yet existing benchmarks on multimodal interactions primarily evaluate response quality, implicitly assuming that the user demand behind a multimodal query has already been correctly identified and understood (Wang et al., 2025; Selvakumar et al., 2025; Lu et al., 2026; Zhao et al., 2026). This assumption is fragile because human expression is inherently underspecified. By the Principle of Least Effort (Zipf, 1949), people tend to avoid unnecessary explicitness, leaving parts of their intended meaning to be recovered from shared multimodal and conversational context. As shown in Fig. 1 (b), expressions such as "the same test" and "this one" require conversation history and visual cues from user gestures to resolve. Conversely, Fig. 1 (e–g) show that demand-like expressions do not necessarily imply the existence of actual user demands. Recovering a demand is therefore a contextual reasoning task that requires integrating multimodal and conversational evidence, not merely transcribing the utterance.

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.

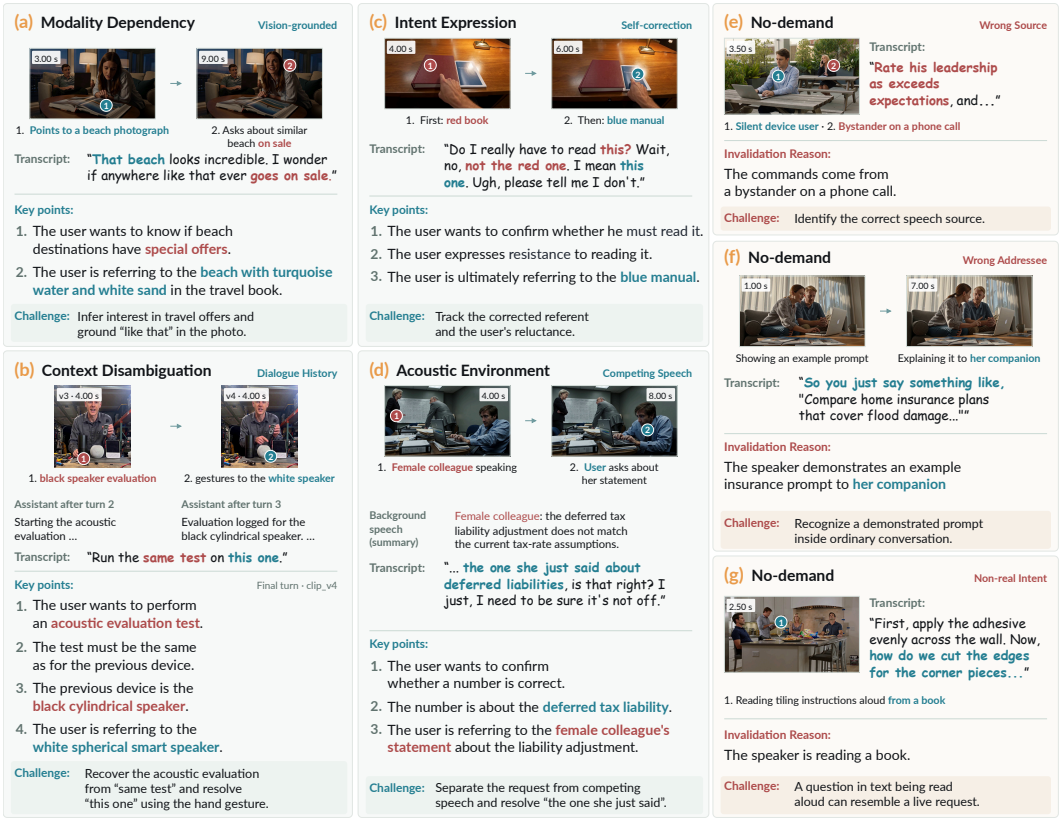


Figure 1: **Representative challenging scenarios in ODU.** Demand-bearing cases (a–d) require integrating visual, acoustic, and conversational context while handling challenging expression forms and acoustic conditions, while no-demand cases (e–g) contain demand-like expressions that should not trigger responses.

To diagnose this capability systematically, we introduce **Omni Demand Understanding (ODU)**, a task which makes demand understanding an explicit prediction target rather than an implicit precondition of response generation. Given an audio or audio–visual interaction, ODU evaluates a model along five dimensions: whether a valid demand is present, when it occurs, the transcript of the demand-bearing speech, the user's intent as resolved from multimodal and conversational context, and the user profile. This diagnosis separates perception from contextual user-intent inference: a model must not only recover the utterance, but also infer the user's intent and the response-relevant information left implicit. Semantic recovery is evaluated against atomic key points, while no-demand scenes test whether request-like signals could trigger the assistant.

To build ODU-Bench, we start from demand-understanding challenges in everyday interaction and organize them into a challenge-driven taxonomy. Guided by sampled taxonomy targets, an agentic pipeline constructs scenarios coarse-to-fine, progressing from demand semantics or invalidation conditions to causal interactions, dialogue, and temporal realization. After challenge-validity and transcript-sufficiency screening, accepted scripts guide the production of synthetic interactions. We also recruited human actors to perform daily-life interactions, providing the benchmark with a complementary source of behaviorally and acoustically realistic data. For both synthetic and human-recorded interactions, demand annotations are reconstructed from the realized media. Every scene and its annotation then undergo careful human verification. Together, these steps yield 2 078 diverse and challenging demand and no-demand scenes. Across a broad panel of native multimodal models, even Gemini 3.1 Pro, the strongest system on our aggregate score, falsely triggers on 40.6% of no-demand scenes. Evidence-channel analysis shows that it recovers 82.9% of key points supported by the spoken request but only 44.7% of those requiring multimodal or dialogue context. This gap confirms that ODU cannot be reduced to transcription: models must select and bind evidence across the interaction to understand the user's intent. In summary, our contributions are as follows:

1. **A new problem formulation.** We establish omni demand understanding as a distinct multimodal contextual inference problem, making it an explicit evaluation target rather than leaving it implicit in response-generation performance.
2. **A challenging benchmark.** We construct a benchmark covering diverse demand and no-demand scenarios, combining taxonomy-guided agentic generation with human-recorded interactions.
3. **A systematic capability gap.** We show that current MLLMs systematically struggle with contextual user-intent inference, particularly when it requires multimodal and conversational evidence beyond the explicit spoken request.

2 Related Work

Multimodal interaction. Research on spoken and multimodal interaction examines response quality, turn-taking, and the coordination of listening and speaking. Full-Duplex-Bench evaluates pause handling, backchanneling, and interruptions (Lin et al., 2025), while MTR-DuplexBench examines multi-round dialogue quality and instruction following (He et al., 2026). Full-Duplex-Bench-v2 further evaluates multi-turn turn-taking and instruction following with an automated examiner (Lin et al., 2026). MultiVox evaluates responses grounded in paralinguistic and visual cues (Selvakumar et al., 2025), and VideoFDB benchmarks nonverbal behavior in full-duplex audio–visual conversations (Mazumdar et al., 2026). OmniMMI studies streaming understanding and multi-turn dependencies (Wang et al., 2025); OmniInteract tests trigger timing, interruptions, and nested exchanges (Lu et al., 2026); and OmniPro emphasizes modality necessity in proactive streaming evaluation (Zhao et al., 2026). ODU complements these benchmarks by evaluating demand understanding separately from response quality and timing.

Multimodal intent and contextual goal understanding. Research on multimodal intent and goal understanding examines how language, behavior, and surrounding context reveal users' intentions. MIntRec2.0 evaluates conversational intent classification using textual, acoustic, and visual cues (Zhang et al., 2024), with substantial textual bias identified in subsequent analyses (Mullick et al., 2025). SIMMC 2.0 studies multimodal disambiguation and coreference in shopping dialogues (Kottur et al., 2021). GUIDE studies assistance needs from GUI activity (Yang et al., 2026); WAGIBench infers unexpressed goals from egocentric context (Veerabadran et al., 2025); and EgoIntrospect evaluates request recovery with the spoken request withheld (Wang et al., 2026). ODU complements these studies with joint evaluation of demand presence and open-ended demand semantics in challenge-driven daily-life human–machine interactions, created through media synthesis and realistic human-recorded interactions. Table 1 compares ODU with representative benchmarks.

Table 1: Benchmark comparison. Demand Presence evaluates whether a demand exists; Intent Prediction Format specifies explicit intent outputs. Input columns denote speech and dialogue history. Visual and Acoustic Cues denote visual and nonlexical acoustic evidence for intent inference.

Benchmark	Demand Presence	Intent Prediction Format	Spoken Utterance Input	Multi-Turn Dialogue Input	Visual Cues	Acoustic Cues
Multimodal interaction benchmarks						
OmniMMI (Wang et al., 2025)	◐	✗	◐	✓	✓	✗
MultiVox (Selvakumar et al., 2025)	✗	✗	✓	✗	✓	✓
OmniInteract (Lu et al., 2026)	◐	✗	✓	✓	✓	✗
Full-Duplex-Bench-v2 (Lin et al., 2026)	◐	✗	✓	✓	✗	✗
Intent and goal understanding benchmarks						
MIntRec2.0 (Zhang et al., 2024)	✗	Label	✓	✓	✓	✓
WAGIBench (Veerabadran et al., 2025)	✗	MCQ / Open-ended	✗	✗	✓	✗
GUIDE (Yang et al., 2026)	✓	MCQ	✗	✗	✓	✗
EgoIntrospect (Wang et al., 2026)	◐	MCQ	◐	✗	✓	✗
ODU	✓	Open-ended	✓	✓	✓	✓

✓: explicitly evaluated in reported settings; ◐: partial or indirect support; ✗: not established. MCQ: multiple-choice question.

3 TASK FORMULATION AND EVALUATION

3.1 PROBLEM SETUP AND PREDICTION TARGETS

A valid *demand* is a current user goal directed to the assistant, expressed explicitly or through the interaction context. ODU evaluates the final user turn in an audio or audio–visual interaction, using the preceding interaction as context. Earlier assistant replies are supplied as text. The task asks **five core questions**: whether a valid demand exists; what the user intends and which contextual information is required to resolve it; when the demand occurs; what the user says; and who expresses it. The first two assess contextual reasoning for demand detection and user-intent inference. The latter three primarily assess perception through temporal localization, transcription, and user profile prediction. Together, they cover the perceptual foundations and contextual reasoning required for demand understanding. The output schema is shown in App. A.

3.2 KEY-POINT-BASED SEMANTIC EVALUATION

We represent demand understanding with two open-ended fields: *structured intent* for the desired outcome and *required context* for the multimodal or dialogue context needed to resolve it. We decompose the reference intent and required context into atomic *key points* covering the requested outcome, its referents, and contextual constraints. M2 measures how much of this reference information a prediction recovers, accepting semantically equivalent wording. Each key point is labeled by evidence source for channel-level diagnosis (§5.3). All key points are grounded in the media and verified by humans. An LLM judge checks each point against the concatenated predicted structured intent and required-context descriptions. We audit key-point recoverability from source annotations and the stability of model conclusions across judges in App. E.2.

3.3 EVALUATION METRICS

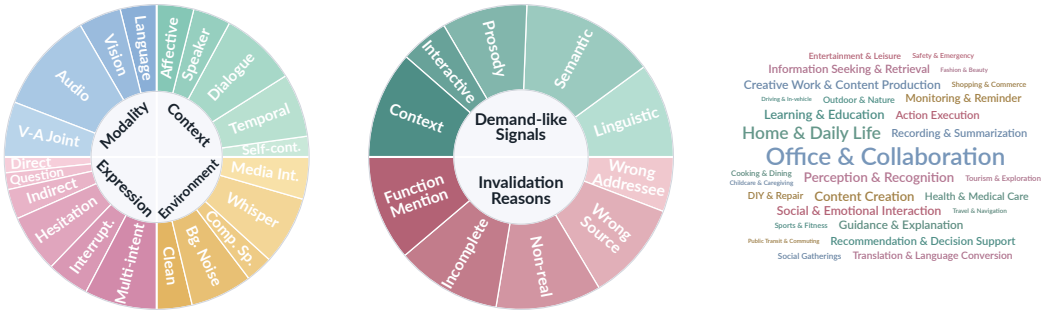
Table 2: Evaluation dimensions and their contribution to the overall score.

	Output measured	**Metric**	**Weight**
M1	demand present	binary macro-F1 over demand and no-demand scenes	0.15
M2	structured intent, required context	key-point hit-rate	0.60
M3	demand span	temporal IoU	0.10
M4	transcript	$\max(0, 1 -$ adaptive CER/WER$)$; CER for CJK-dominant references and WER otherwise	0.10
M5	user profile	mean per-field accuracy	0.05

Table 2 summarizes M1–M5. M1 uses demand and no-demand scenes; M2–M5 use only demand-bearing scenes. FTR is the fraction of no-demand scenes that trigger the assistant. We use a task-priority weighted average: M2 (0.60) receives the largest weight because recovering intent and required context is the central goal. M1 (0.15) comes next because distinguishing demands from non-demands is essential to appropriate engagement. M3 and M4 (0.10 each) receive equal, smaller weights as supporting perception measures for localization and transcription. M5 (0.05) is weighted least because user profile is auxiliary to demand semantics. We report all dimensions separately alongside this aggregate. App. A gives full definitions and aggregation details.

4 BENCHMARK CONSTRUCTION

Fig. 3 summarizes ODU-Bench construction for audio-only and audio–visual interactions: a challenge-driven taxonomy and seed corpus (§4.1), agentic scenario generation (§4.2), and media-grounded annotation with quality control (§4.3). App. C details construction and human verification procedures.



(a) Distribution of positive samples by taxonomy axis.

(b) Distribution of negative samples by signal type and invalidation pattern.

(c) Coverage across task and scenario categories.

Figure 2: **Taxonomy coverage in ODU-Bench.** (a) and (b) show positive and negative sample distributions across their taxonomy axes; segment angles are proportional to sample counts within each axis. (c) shows task and scenario coverage, with word size reflecting frequency.

4.1 Challenge-driven taxonomy and seed corpus

The core challenge of user demand understanding lies in how demands are expressed and how the surrounding context shapes their interpretation. Real-world demands may depend on visual or acoustic context, remain ambiguous without prior interaction, be expressed implicitly, or occur under distracting environmental conditions. In contrast, task semantics and scenario domains primarily characterize what users want to accomplish and where interactions occur, rather than what makes the demands difficult to understand. We therefore organize scenario generation around sources of demand-understanding difficulty, while separately characterizing task and domain coverage.

We define a six-axis taxonomy for positive-demand scenes: four independently composable challenge axes for generation and two descriptive axes for task and domain coverage. No-demand scenes use a separate taxonomy of demand-like signal types and invalidation reasons. Fig. 2 shows the taxonomy and resulting distribution; App. B provides the definitions.

We combine seeds from a corpus of everyday contexts and user needs with sampled taxonomy targets to vary the setting of each challenge. This separates *challenge coverage* from *scenario diversity*.

4.2 Agentic scenario generation

Hierarchical script generation. Starting from taxonomy targets and scenario seeds (§4.1), Qwen3.7-Max (Qwen Team, 2026b) builds a *scene skeleton* specifying characters, objects, and the setting. The main demand-generation path specifies *user intent* separately from its expression: it builds supporting context and dialogue around the intent before generating the user's utterance. No-demand scenes instead instantiate demand-like signals and their invalidating conditions. We assemble these components into a *timeline* for one or more clips.

Script validity review. The LLM reviewer checks *plausibility, consistency, taxonomy alignment, and challenge validity*, including causal coherence and demand recoverability from scripted evidence. The *transcript-only discriminator* receives only the scripted user utterance, with dialogue history, speaker information, and media evidence withheld. For scenarios designed to require contextual evidence, we reject text-sufficient scripts: positives whose intent is recoverable from the utterance alone and negatives whose no-demand status is textually obvious.

Media realization from scripts. Accepted scripts follow the two branches in Fig. 3. For *synthesized media*, clips are rendered sequentially, conditioning later clips on earlier videos to preserve character and environmental continuity; audio-only scenes retain the audio track. For *human-recorded media*, scripts serve as performance guides, preserving taxonomy targets and interaction structure while introducing real actors, environments, microphones, timing, and speech variation. Both branches undergo automatic review and human inspection for perceptual quality, scenario fidelity, and cross-clip coherence before *media-grounded annotation and quality control* (§4.3).

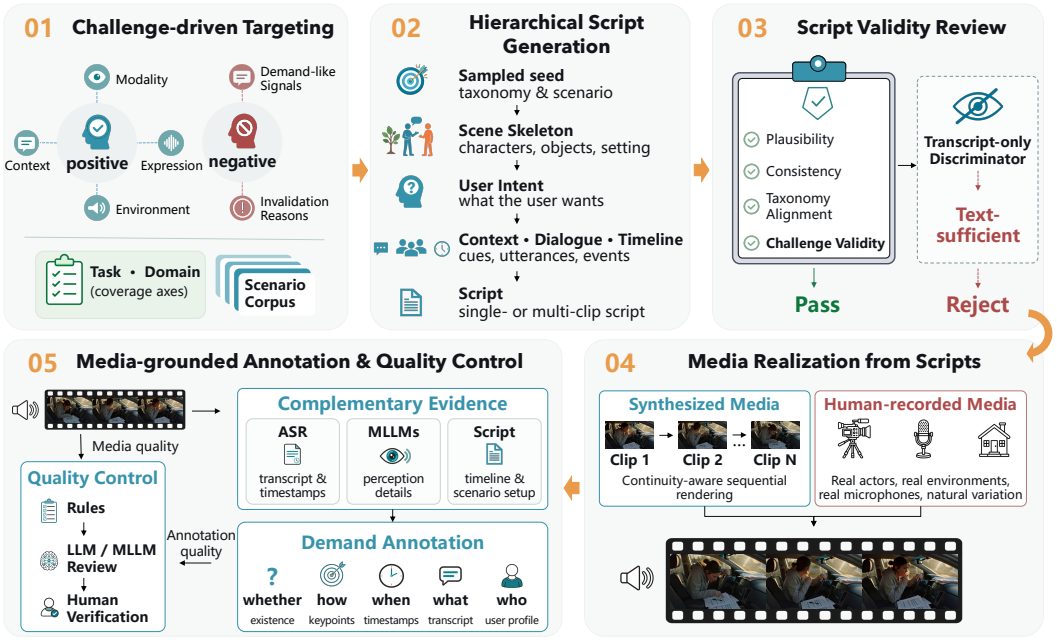


Figure 3: **Overview of the ODU-Bench construction pipeline.** Challenge-driven targets are expanded into coarse-to-fine scripts, screened for plausibility and challenge validity, and realized as synthesized or human-recorded interactions. Annotations are then reconstructed from complementary media evidence and verified through automatic checks and human review.

4.3 Media-grounded annotation and quality control

Rendered and human-recorded media may differ from scripts in scene details, speech, timing, and demand boundaries. The initial annotation draft combines word-level ASR (An et al., 2025) with media observations from MLLMs (Qwen Team, 2026c; Google DeepMind, 2026), without the script's intended annotation. We compare the draft with script-based expectations and, when significant discrepancies arise, use targeted MLLM observations of the media to guide annotation refinement.

Rule-based checks and LLM/MLLM review assess consistency between scripts and media, transcript and timestamp alignment, schema validity, agreement between demand labels and segments, context support, and key-point grounding. Unsupported key points are removed or rewritten. Human reviewers check machine-produced fields against the realized interaction, reject low-quality samples, and retain generated scenes only when all applicable fields pass review; experts correct annotation errors in human-recorded interactions. The released dataset contains 1 801 synthetic scenarios and 277 human-recorded interactions performed by 30 actors.

5 Experiments

5.1 Experimental setup

We evaluate 14 native configurations (Table 3): 12 on audio–visual scenes and 14 on audio-only scenes, using the same task instruction and modality-specific output schema.
GPT-Realtime-2 (OpenAI, 2026b) and Kimi-Audio (KimiTeam et al., 2025) accept audio only; all other systems are evaluated on both modalities, with results reported separately. For multi-turn interactions, only the final user clip is scored (§3.1). As a strong text-only baseline,
GPT-5.4 (OpenAI, 2026a) receives ASR transcripts, sentence timings, and prior-turn assistant reply text (App. D.2). It remains unranked because its input interface differs from that of the native systems. We use Qwen3.6-Flash (Qwen Team, 2026a) as the LLM judge. Re-judging the audited model subset with alternative judges preserves its key-point coverage ranking (App. E.2).

Table 3: **Evaluation results on the ODU task.** Audio–visual (AV) and audio-only (AO) results are reported separately as percentages. M1–M5 measure detection, key-point coverage, localization, transcription, and user profile; Avg. uses weights 0.15/0.60/0.10/0.10/0.05 over M1–M5, and FTR is the no-demand false-trigger rate (lower is better). Size denotes the model parameter count; for MoE models, "A" indicates the number of active parameters. The text-only system receives ASR transcripts and sentence-level timestamps but no media.

Model	Size	Audio–visual (%)							Audio-only (%)						
		Avg.	M1 Detect	M2 Keypt.	M3 Locate	M4 Trans.	M5 Profile	FTR Neg.	Avg.	M1 Detect	M2 Keypt.	M3 Locate	M4 Trans.	M5 Profile	FTR Neg.
Closed-source Models															
Gemini 3.1 Pro	–	**72.6**	83.0	63.4	81.9	**91.3**	**95.7**	40.6	75.4	87.3	70.1	61.5	91.7	**97.6**	31.7
Gemini 3.7 Flash	–	69.6	**83.9**	60.2	78.7	85.9	88.3	**27.5**	72.9	**89.6**	66.1	62.7	88.4	94.1	**16.0**
Gemini 3.5 Flash Lite	–	57.4	63.9	49.9	68.1	71.5	79.3	55.6	62.2	74.5	56.4	56.7	74.3	82.4	38.5
Qwen3.5-Omni-Plus	–	69.6	63.9	**65.8**	75.7	83.3	93.5	73.8	74.7	76.2	71.5	66.1	89.3	95.8	55.3
Seed 2.0 Lite	–	67.9	58.1	64.1	**82.8**	77.7	92.5	82.2	**77.3**	60.8	**78.0**	**71.4**	**93.2**	97.5	80.7
GPT-Realtime-2	–	–	–	–	–	–	–	–	63.8	71.0	63.4	34.3	81.4	71.2	63.4
Open-source Models															
Qwen3-Omni-Think	30B-A3B	59.9	58.2	55.6	47.0	85.2	92.1	82.9	67.4	73.7	63.0	50.5	88.1	94.6	60.9
Qwen3-Omni-Instruct	30B-A3B	55.2	50.6	53.1	23.9	87.1	92.8	93.4	61.6	61.0	60.9	20.5	91.0	95.2	80.7
Qwen2.5-Omni	7B	45.2	53.8	42.1	5.8	74.2	78.2	87.8	48.0	67.1	45.7	5.5	69.4	60.7	66.5
Ming-Flash-Omni 2.0	104B-A6B	52.6	50.1	49.5	29.6	83.1	82.3	93.7	52.6	63.0	50.5	7.3	86.1	69.6	75.8
MiniCPM-o 4.5	9B	39.1	54.5	35.5	1.7	60.3	68.1	85.3	49.1	65.1	47.9	6.2	61.5	75.9	74.5
Nemotron 3 Nano Omni	30B-A3B	34.9	55.8	26.8	35.9	34.3	69.6	72.0	40.1	59.4	33.8	39.3	37.6	64.1	72.0
video-SALMONN 2+	7B	14.2	50.3	6.7	2.9	12.3	23.4	89.1	23.8	46.4	18.4	10.0	26.3	42.6	97.5
Kimi-Audio	7B	–	–	–	–	–	–	–	52.6	58.4	51.1	17.2	71.8	85.0	82.0
Text-only System (no native media input)															
GPT-5.4 + ASR transcript	–	59.0	60.6	55.5	82.3	81.9	2.6	73.4	71.5	78.9	68.3	86.1	85.1	30.6	40.4

Note. Best and second-best values are **bold** and underlined.

5.2 Overall results

ODU remains challenging even for the strongest models. Table 3 shows that the best native Avg. scores reach only 72.6% on AV scenes (Gemini 3.1 Pro) and 77.3% on AO scenes (Seed 2.0 Lite). The highest AV M2 is just 65.8%, leaving substantial headroom in the primary semantic dimension. Open-source models lag further: even their strongest model, Qwen3-Omni-Think, reaches only 59.9% AV and 67.4% AO Avg.

Strong perception does not imply strong contextual reasoning. On AV scenes, Gemini 3.1 Pro achieves 81.9% M3, 91.3% M4, and 95.7% M5, yet its M2 is only 63.4% and its FTR reaches 40.6%. High localization, transcription, and user-profile scores thus do not ensure contextual demand understanding. Further analyses (§5.3, §5.4) and qualitative examples (App. E.5) reveal missed contextual information and false triggers from request-like speech or source/addressee mismatches.

Strong demand recovery does not ensure reliable engagement. Seed 2.0 Lite achieves 78.0% M2 on audio-only scenes but triggers on 80.7% of no-demand scenes. Gemini 3.7 Flash has lower M2 (66.1%) and much lower FTR (16.0%). Reliable demand understanding therefore requires both interpreting what users want and deciding whether to respond.

Models targeting efficient or real-time interaction lag in demand understanding. Gemini 3.5 Flash Lite, GPT-Realtime-2, and MiniCPM-o 4.5 achieve AO Avg. scores of only 62.2%, 63.8%, and 49.1%, respectively, below both Gemini 3.1 Pro (75.4%) and Qwen3-Omni-Think (67.4%). This gap raises a practical concern for interactive systems that rely on real-time models: weak demand understanding can cause responses to miss the user's intent.

5.3 Contextual demand recovery

To examine which information is missed, we compare Gemini 3.1 Pro, Seed 2.0 Lite, Qwen3.5-Omni-Plus, and the GPT-5.4 text-only baseline on the audio–visual demand scenes. Figure 4 reports

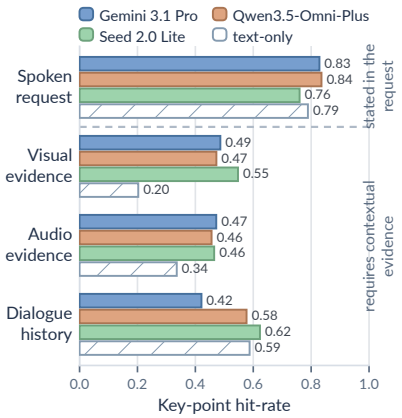


Figure 4: **Key-point hit-rate by evidence channel** on AV demand scenes. Rates are point-weighted; hatched bars denote the GPT-5.4 text-only baseline (App. E.1).

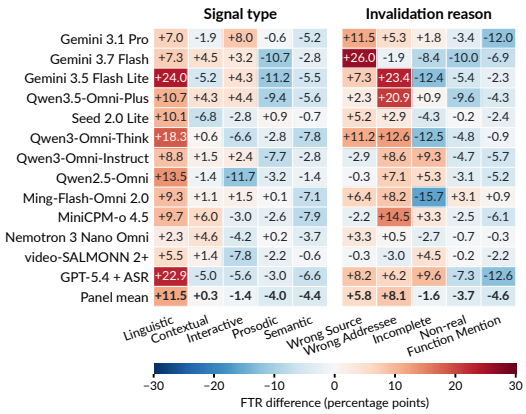


Figure 5: **False-trigger patterns on pooled AV/AO negatives.** Cells show subtype FTR minus the same system's overall FTR on labeled negatives, in percentage points. Dedicated audio-only systems are outside this panel. The final row is the unweighted mean across displayed systems.

point-weighted key-point hit rates by annotated evidence source: spoken request, visual evidence, acoustic evidence, and dialogue history.

Models recover multimodal and conversational context less reliably than spoken requests. Native spoken-request hit-rates are 76.0–83.6%, comparable to the text-only baseline's 78.9% and consistent with the strong transcription scores in Table 3. Yet all three native systems score below 60% on visual and acoustic key points, highlighting weak recovery of multimodal context. Dialogue-history hit-rates are also lower (42.1–62.3% for native systems). The text-only baseline receives prior-turn assistant reply text and reaches 58.7%. The text-only baseline also recovers some visual and acoustic key points, possibly by leveraging commonsense knowledge together with scene details inferred from transcripts and dialogue history. **These results show that accurate speech transcription alone is insufficient for strong ODU performance.**

5.4 Error patterns across the taxonomy

We group whole scenes by taxonomy category to identify recurring failure patterns. Figure 5 reports subtype FTR minus each model's overall FTR on the combined, labeled AV and AO negative scenes, with separate groupings by signal type and invalidation reason. For positives, we compare category M2 with each model's mean over the common positive cohort, separately for AV and AO (App. E.3).

Models falsely trigger on negative scenes with request-like wording or source/addressee mismatches. Linguistic signals such as questions, imperatives, and complaints yield the highest average FTR, exceeding every model's mean with an average gap of 11.5 percentage points, whereas semantic signals that merely mention assistant functions fall below every model's mean. Wrong Source and Wrong Addressee are the hardest invalidation categories on average; even Gemini 3.7 Flash, with the lowest FTR in Table 3, exceeds its own mean by 26.0 percentage points on Wrong Source. In these scenes, models frequently interpret request-like speech as a demand directed to the assistant, even when the surrounding context indicates that the speech comes from media playback or background speakers (Wrong Source), or is addressed to someone else (Wrong Addressee).

Models struggle to recover indirect requests on positive scenes. Indirect or descriptive requests score below every model's mean M2, with average differences of -7.2 points on AV scenes and -8.2 points on AO scenes. Models can mistake speech for a request to the assistant, yet struggle to understand genuine needs that users do not state directly.

Table 4: Model performance (%) on synthetic (Syn.) and human-recorded (Rec.) audio–visual subsets. Δ = Rec. − Syn. in percentage points.

Metric	Gemini 3.1 Pro			Seed 2.0 Lite			Qwen3.5-Omni-Plus			Qwen3-Omni-Think			Ming-Flash-Omni 2.0		
	Syn.	Rec.	Δ	Syn.	Rec.	Δ	Syn.	Rec.	Δ	Syn.	Rec.	Δ	Syn.	Rec.	Δ
Avg.	71.9	74.2	+2.4	67.5	67.6	+0.1	68.5	65.7	−2.8	61.3	53.6	−7.6	54.2	49.3	−4.8
M1 Detect	83.3	86.0	+2.7	58.2	55.9	−2.2	60.7	53.7	−6.9	60.9	48.6	−12.3	49.2	43.5	−5.6
M2 Keypt.	62.0	65.1	+3.2	63.9	63.7	−0.2	64.9	64.2	−0.7	56.4	49.5	−6.9	50.8	47.6	−3.2
M3 Locate	81.8	84.9	+3.1	82.6	82.9	+0.3	76.3	73.1	−3.3	45.9	49.8	+3.9	31.0	37.5	+6.5
M4 Trans.	92.3	88.3	−4.1	76.3	79.8	+3.5	81.9	71.6	−10.3	90.7	71.3	−19.5	89.5	62.3	−27.3
M5 Profile	95.3	98.3	+3.1	91.5	94.9	+3.4	93.5	93.3	−0.2	92.8	90.5	−2.2	84.6	85.1	+0.5
FTR	40.9	34.7	−6.3	81.9	84.0	+2.1	79.0	86.7	+7.6	80.9	92.0	+11.1	95.2	98.7	+3.4

5.5 Comparison with human-recorded interactions

To assess demand understanding beyond synthesized interactions, we compare synthetic and human-recorded audio–visual subsets (Table 4).

Leading MLLMs retain their aggregate performance on human-recorded interactions. Gemini 3.1 Pro and Seed 2.0 Lite have similar observed Avg. scores across the two subsets. Qwen3-Omni-Think and Ming-Flash-Omni 2.0 score lower on the human-recorded subset, with Avg. differences of -7.6 and -4.8 points, respectively.

Human-recorded speech exposes transcription weaknesses. Qwen3.5-Omni-Plus and Qwen3-Omni-Think show their largest declines in M4. Ming-Flash-Omni 2.0 has the steepest transcription decline in the table: -27.3 points in M4, versus -3.2 in M2. These declines may partly reflect more complex environmental noise and accented speech in realistic interactions.

5.6 Downstream response quality

We further examine whether correct demand understanding translates into better assistant responses. For each of 200 sampled audio–visual scenes, Gemini 3.1 Pro generated two responses: one from the media and dialogue history alone, and one from the same inputs plus a reference demand annotation. The annotation described what the user wanted and the relevant context, or stated that no demand was present. Both addressed the final user turn under identical generation settings (App. E.4).

In a blinded human A/B test with randomized response order, the evaluator chose a preferred response or tie based on correctness, relevance, contextual grounding, and appropriate silence. Ties accounted for 40.0%; among decisive judgments, 85.8% favored replies *with demand annotation*. Thus, **providing correct demand information substantially improves downstream response quality in our evaluation**, underscoring the practical importance of demand understanding for effective conversational assistants.

6 Conclusion

We introduced Omni Demand Understanding (ODU) as a distinct task for identifying valid user demands and inferring intent from multimodal and conversational context. ODU-Bench combines taxonomy-guided agentic generation with human-recorded interactions and uses human-verified annotations to evaluate demand understanding along five complementary dimensions. Experiments with 14 native MLLMs show that strong perceptual performance does not ensure accurate contextual user-intent reasoning. In our evidence-channel analysis, the three native systems evaluated on audio–visual demand scenes recover multimodal and conversational context less reliably than spoken requests, highlighting the difficulty of contextual reasoning. Models also frequently mistake request-like speech for a demand to the assistant, even when it comes from media playback or is addressed to someone else. A blinded human A/B study further shows that providing reference demand annotations improves response quality, underscoring the practical value of correct demand understanding. We hope ODU will make demand understanding a central focus of multimodal interaction research and drive progress toward assistants that infer user intent from multimodal and conversational context before deciding whether and how to respond.

AI USE STATEMENT

We used generative AI tools to assist in designing the benchmark taxonomy. Generative AI tools are integral components of our pipeline for data synthesis, annotation, and quality control (§4; App. C). We also use an LLM judge for key-point-based semantic evaluation (§3.2; App. A). Additionally, we used AI writing assistants to improve the writing of this paper. We did not use generative AI to develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for proving mathematical claims; assist in writing mathematical proofs; propose or refine hypotheses; or design research methodology or experiments. All benchmark media and annotations underwent rigorous human review, and human reviewers determined inclusion in the released benchmark. We reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Human-recorded interactions were performed by paid actors. Before release, all participating actors signed informed-consent forms authorizing use of their recording data, including their faces and voices, for academic evaluation and public release with the benchmark. To protect actor privacy, the released human-recorded interactions may be used solely to evaluate systems on this benchmark; all other academic and commercial uses are prohibited. Generated scenes contain no real individuals. ODU-Bench is provided solely for academic research and model evaluation, subject to the above restrictions on human-recorded interactions. Commercial use of the dataset or its images in products, services, or other profit-making activities is prohibited. Images in the dataset may depict identifiable individuals; copyright and related rights remain with their respective rights holders. The release does not authorize sublicensing, commercial exploitation, or the creation of derivative works from these images. Users must comply with the release terms, the scope of participant consent, and applicable personal information protection and portrait rights laws and regulations.

REPRODUCIBILITY STATEMENT

The main paper defines the task, construction procedure, and evaluation framework (§3; §4). The appendix specifies the prediction schema, matching, metrics, aggregation, and a worked scoring example (App. A), the release composition and taxonomy (App. B), and the construction and human-verification procedures (App. C). Experimental input preparation, the transcript-only baseline, and the judge configuration are documented in App. D; supplementary analyses and judge stability are reported in App. E. Evaluation instructions are provided in App. F. Benchmark media, annotations, evaluation code, and configurations provide the inputs and tools for evaluating ODU.

REFERENCES

Inclusion AI, Bowen Ma, Cheng Zou, ChengKun Du, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Chenyu Lian, Chengxiang Fan, Dandan Zheng, et al. Ming-Flash-Omni: A sparse, unified architecture for multimodal perception and generation. *arXiv preprint arXiv:2510.24821*, 2025.

Keyu An, Yanni Chen, Zhigao Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Bo Gong, Xiangang Li, Yabin Li, Ying Liu, Xiang Lv, Yunjie Ji, Yiheng Jiang, Bin Ma, Haoneng Luo, Chongjia Ni, Zexu Pan, Yiping Peng, Zhendong Peng, Peiyao Wang, Hao Wang, Haoxu Wang, Wen Wang, Wupeng Wang, Yuzhong Wu, Biao Tian, Zhentao Tan, Nan Yang, Bin Yuan, Jieping Ye, Jixing Yu, Qinglin Zhang, Kun Zou, Han Zhao, Shengkui Zhao, Jingren Zhou, and Yanqiao Zhu. Fun-ASR technical report, 2025. URL https://arxiv.org/abs/2509.12508.

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, 2026. URL https://arxiv.org/abs/2604.27393.

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, et al. Nemotron 3 nano omni: efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

Google DeepMind. Gemini 3.1 pro. https://deepmind.google/models/gemini/pro/, 2026.

Zhang He, Wenqian Cui, Haoning Xu, Xiao-Hui Li, Lei Zhu, Haoli Bai, Ma Shaohua, and Irwin King. MTR-DuplexBench: Towards a comprehensive evaluation of multi-round conversations for full-duplex speech language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5334–5351, 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.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-Audio technical report, 2025. URL https://arxiv.org/abs/2504.18425.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4903–4912, 2021.

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

Guan-Ting Lin, Shih-Yun Shan Kuan, Jiatong Shi, Kai-Wei Chang, Siddhant Arora, Shinji Watanabe, and Hung-yi Lee. Full-Duplex-Bench-v2: A multi-turn evaluation framework for duplex dialogue systems with an automated examiner. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 27–36. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-short.4. URL https://aclanthology.org/2026.acl-short.4/.

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

Ankan Mullick, Saransh Sharma, Abhik Jana, and Pawan Goyal. Text takes over: A study of modality bias in multimodal intent detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24039–24069, 2025.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026a.

OpenAI. Advancing voice intelligence with new models in the api. https://openai.com/index/advancing-voice-intelligence-with-new-models-in-the-api/, 2026b.

Qwen Team. Qwen3.6-35b-a3b: Agentic coding power, now open to all, 2026a. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Qwen Team. Qwen3.7: The agent frontier, 2026b. URL https://qwen.ai/blog?id=qwen3.7.

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

Bytedance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

Ramaneswaran Selvakumar, Ashish Seth, Nishit Anand, Utkarsh Tyagi, Sonal Kumar, Sreyan Ghosh, and Dinesh Manocha. MULTIVOX: A benchmark for evaluating voice assistants for multimodal interactions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28469–28481, 2025.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Caption-Enhanced Audio-Visual Large Language Models. *arXiv preprint arXiv:2506.15220*, 2025.

Vijay Veerabadran, Fanyi Xiao, Nitin Kamra, Pedro Matias, Joy Chen, Caley Drooff, Brett Roads, Riley J Williams, Ethan Henderson, Xuanyi Zhao, et al. Benchmarking egocentric multimodal goal inference for assistive wearable agents. *Advances in Neural Information Processing Systems*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18925–18935, 2025.

Zeyu Wang, Chang Liu, Eduardus Tjitrahardja, Yuntao Wang, Borislav Pavlov, Fangfei Gou, Jose Manuel Davila, Dai Shi, Ran Xu, Yue Pan, et al. EgoIntrospect: An egocentric dataset and benchmark for user-centric internal state reasoning. *arXiv preprint arXiv:2605.17262*, 2026.

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, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Saelyne Yang, Jaesang Yu, Yi-Hao Peng, Kevin Qinghong Lin, Jae Won Cho, Yale Song, and Juho Kim. GUIDE: A benchmark for understanding and assisting users in open-ended GUI tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13017–13027, 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 *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nY9nITZQjc.

Ruixiang Zhao, Jie Yang, Zijie Xin, Tianyi Wang, Fengyun Rao, Jing LYU, and Xirong Li. OmniPro: A comprehensive benchmark for omni-proactive streaming video understanding. *arXiv preprint arXiv:2605.18577*, 2026.

George Kingsley Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, 1949.

APPENDIX CONTENTS

A Task specification and evaluation **14**

A.1 Prediction schema . . . 14

A.2 Segment matching . . . 14

A.3 Metric definitions . . . 14

A.4 Aggregation and key-point accounting . . . 15

A.5 Model output and scoring walkthrough . . . 16

B Benchmark composition and taxonomy **17**

B.1 Dataset statistics . . . 17

B.2 Positive-demand scenes: six axes . . . 17

B.3 No-demand scenes: signal type and invalidation reason . . . 19

C Construction and human verification **20**

C.1 Target selection and seed sampling . . . 20

C.2 Script generation and review . . . 20

C.3 Media production and annotation . . . 21

C.4 Human verification . . . 22

D Experimental setup **25**

D.1 Input preparation and evaluation . . . 25

D.2 Text-only baseline . . . 25

D.3 Semantic scoring . . . 26

E Supplementary experiments and diagnostics **26**

E.1 Request and context key-point coverage . . . 26

E.2 LLM judge stability . . . 28

E.3 Positive-demand challenge breakdown . . . 29

E.4 Blinded response comparison . . . 31

E.5 Qualitative error analysis . . . 33

F Evaluation prompts **35**

F.1 Shared prediction task . . . 35

F.2 Audio-only variant . . . 37

F.3 Text-only input structure . . . 37

F.4 Key-point judge . . . 38

A TASK SPECIFICATION AND EVALUATION

A.1 PREDICTION SCHEMA

Only the final user clip is annotated, and timestamps are measured from its start. A prediction contains a `has_demand` flag and, when it is true, a list of demand segments. Each segment carries a time span, a transcript, a structured intent, a required-context list, and closed-set user profile fields. The annotation reference also includes a frozen key-point set and a necessity note for each required-context item. Fig. 6 gives the schema, and App. F.1 gives the system prompts for evaluation.

```
{
  "has_demand": bool,
  "segments": [{
    "start_ms": int, "end_ms": int,            // demand span, milliseconds
    "transcript":        str,                  // the demand-bearing speech
    "structured_intent": str,                  // what the user wants done
    "required_context": [                      // what the words left out
      {"type": "visual" | "audio" | "dialogue_history" | "user_activity"
             | "user_identity" | "user_emotion" | "scene_condition",
       "description": str}
    ],
    "user_in_frame": bool, "user_gender": str, "user_age_group": str
  }]
}
```

Figure 6: The prediction schema. A model emits `has_demand` and, when it is true, one segment per continuous demand-bearing region. Audio-only predictions omit visual context and `user_in_frame`.

A.2 SEGMENT MATCHING

A scene may contain multiple demand segments, so predicted segments are first aligned with reference segments before computing M2–M5. We prioritize normalized-transcript similarity, which is more robust to boundary drift, and fall back to temporal IoU for remaining segments. Let N remove punctuation, control characters, and whitespace and apply case folding. For predicted and reference transcripts p and g, similarity is

$$\mathrm{sim}(p,g) = \begin{cases} \dfrac{2\sum_{b\in\mathcal{B}(N(p),N(g))}|b|}{|N(p)|+|N(g)|}, & |N(p)|>0 \ \wedge\ |N(g)|>0, \\ 0, & \text{otherwise.} \end{cases}$$

Here $\mathcal{B}$ is the matching-block list returned by Python's `difflib.SequenceMatcher` with `autojunk=False`, with prediction and reference supplied in that order. We additionally resolve common segmentation mismatches by allowing one prediction to cover multiple reference segments and by merging multiple predicted fragments that correspond to the same reference segment. This prevents over-splitting or over-merging from being penalized as a semantic error. Unmatched reference segments are retained and scored as missing in the corresponding metrics. Matching thresholds, tie handling, and fragment-merging rules are specified in the released evaluation code.

A.3 METRIC DEFINITIONS

Let $\mathcal{D}$ be the evaluation set, $\mathcal{D}^+$ its demand-bearing subset, and G_x the reference segments of scene x. We distinguish scene scores $s_m(x)$ (M2–M5) from dataset scores M_m (M1–M5).

M1: Demand detection. For each scene $x \in \mathcal{D}$, let y_x and $\hat{y}_x$ be the reference and predicted demand-presence labels. With one-vs-rest counts TP_c, FP_c, and FN_c for each class $c \in \{0,1\}$,

$$M_1 = \frac{1}{2}\sum_{c\in\{0,1\}} \frac{2\mathrm{TP}_c}{2\mathrm{TP}_c+\mathrm{FP}_c+\mathrm{FN}_c}, \qquad \mathrm{FTR} = \frac{|\{x\in\mathcal{D}:\ y_x=0,\ \hat{y}_x=1\}|}{|\{x\in\mathcal{D}:\ y_x=0\}|}. \tag{1}$$

The false-trigger rate is reported separately and does not enter Avg.

M2: Key-point coverage. Let K_g be the frozen key-point set for reference segment g. The LLM judge assigns a hit decision $h_{gk} \in \{0, 1\}$ to each point k using the concatenated predicted intent and required-context descriptions. We count each distinct point once, regardless of tier labels, and average the per-segment hit rates equally within each scene:

$$s_2(x) = \frac{1}{|G_x^K|} \sum_{g \in G_x^K} \frac{1}{|K_g|} \sum_{k \in K_g} h_{gk}, \qquad G_x^K = \{g \in G_x : |K_g| > 0\}, \tag{2}$$

where $h_{gk} = 0$ for every point in an unmatched reference segment. Thus a missed demand cannot improve coverage by removing its points from the denominator.

M3: Segment localization. Write $m(g)$ for the prediction matched to reference segment g and set its IoU to zero when no match exists. The per-scene score averages IoU over all reference segments:

$$s_3(x) = \frac{1}{|G_x|} \sum_{g \in G_x} \mathrm{IoU}(m(g), g)\,. \tag{3}$$

Span precision, recall, F1, and the start- and end-time mean absolute errors are retained as diagnostics but do not enter Avg.

M4: Transcript quality. Let E_g be the Levenshtein edit count and L_g the reference length after case folding and whitespace and punctuation normalization. Characters are used for a CJK-dominant reference segment and words otherwise; an unmatched reference is compared with an empty hypothesis. Let $E = \sum_g E_g$, $L = \sum_g L_g$, and H_g be the normalized hypothesis length, with $H_g = 0$ for an unmatched reference. We pool edit counts and reference lengths across segments, so longer transcripts carry more weight within a scene:

$$s_4(x) = \begin{cases} \max(0, 1 - E/L), & L > 0, \\ 1 - \dfrac{1}{|G_x|} \sum_{g \in G_x} \mathbf{1}[H_g > 0], & L = 0,\ |G_x| > 0, \\ \text{undefined}, & |G_x| = 0. \end{cases} \tag{4}$$

When total reference length is zero, we average per-segment empty-reference error rates: zero for an empty hypothesis and one otherwise. Scenes without reference segments have no defined M4 score.

M5: User profile. Let F_x contain user-in-frame, gender, and age group for audio–visual scenes, and gender and age group for audio-only scenes. Write u_{gf} and $\hat{u}_{gf}$ for the reference and predicted values of field f for segment g. The score averages accuracy over fields and reference segments, with unmatched segments scored as incorrect:

$$s_5(x) = \frac{1}{|F_x|} \sum_{f \in F_x} \frac{1}{|G_x|} \sum_{g \in G_x} \mathbf{1}[\hat{u}_{gf} = u_{gf}]\,. \tag{5}$$

A.4 Aggregation and key-point accounting

For $m \in \{2, 3, 4, 5\}$, let $\mathcal{D}_m^+ \subseteq \mathcal{D}^+$ contain the positive scenes on which the dimension score is defined. Each such scene receives equal weight in the reported score, $M_m = |\mathcal{D}_m^+|^{-1} \sum_{x \in \mathcal{D}_m^+} s_m(x)$. M1 is defined by Eq. 1 over both positive and no-demand scenes. Let A be the set of available dimensions for a model. Avg. is

$$\mathrm{Avg} = \frac{\sum_{m \in A} w_m M_m}{\sum_{m \in A} w_m}, \qquad (w_1, w_2, w_3, w_4, w_5) = (0.15, 0.60, 0.10, 0.10, 0.05). \tag{6}$$

For no-demand scenes, only M1 and the separately reported false-trigger rate are defined.

Tier labels indicate whether a point belongs to demand semantics (T1), contextual grounding (T2), or both. M2 counts their union once. Separately, the source label assigns each attributed point to an evidence channel: the spoken request, visual evidence, audio evidence, dialogue history, or other context. Points with no assigned channel remain in the overall reference denominator but are excluded from channel-specific rates. App. E.1 gives the channel analysis and its population accounting.

A.5 MODEL OUTPUT AND SCORING WALKTHROUGH

The following released example shows the reference annotation, a stored model prediction, the judge input, and the resulting scene-level scores. The judge configuration is specified in App. D.3; the final benchmark aggregation follows Eq. 6.

Reference annotation. Scene `avp_gen_000139` contains one reference segment at 3510–5390 ms. The user says *"Hey, agent, mute that sound."* The structured intent is *"The user wants to mute the electronic notification sound currently playing"*. Its required context is:

`audio` The device in the close foreground emits a short notification chime

`visual` At the bottom edge of the frame, a hand holds a glowing smartphone in landscape orientation, which is the source device emitting the notification chime

The reference user profile is `user_in_frame=true, user_gender=male, user_age_group=adult`.

Model prediction and judge input. Seed 2.0 Lite predicts a demand at 3600–5500 ms, with transcript *"Hey agent, mute that sound."* The scorer concatenates its predicted intent and required-context descriptions into the following candidate text: *"The user wants the AI agent to mute the ongoing notification sound from his mobile phone. | audio: A repeated beeping notification sound from the user's phone is playing."*

The judge receives this candidate text together with the reference key points listed below. The reference transcript and context-necessity notes are not appended to the candidate text.

The predicted user profile is `user_in_frame=false, user_gender=male, user_age_group=adult`.

Stored key-point verdicts.

	Evidence channel	Verdict	Key point
kp01	`intent`	hit	The user wants to mute the electronic notification sound currently playing
kp02	`required_context.audio`	hit	The sound to be muted is the notification chime from the nearby device
kp03	`required_context.visual`	miss	The sound source is the smartphone held by a hand at the bottom of the frame

The stored judge marks 2 of the 3 points as hits, giving this scene an M2 score of $2/3 = 0.67$. Although the prediction names a mobile phone, the stored judge marks kp03 as a miss because it omits the hand-held device and its position at the bottom of the frame. These verdicts concern coverage of the reference points in the predicted text; they do not directly measure perception.

Scene-level scores. The temporal IoU is 0.90, with start- and end-time errors of 90 ms and 110 ms. The normalized transcripts match exactly. Gender and age group are correct, while `user_in_frame` is incorrect, yielding two correct profile fields out of three. The scene-level scores are:

	Dimension	Score
M2	Key-point Coverage	0.67
M3	Segment Localization	0.90
M4	Transcript Quality	1.00
M5	User profile	0.67

Demand detection is correct for this scene. The reported M1, however, is macro-F1 computed across demand and no-demand scenes. Consequently, the benchmark Avg. is computed from the dataset-level dimension scores in Eq. 6, rather than by treating this scene's detection correctness as M1.

B BENCHMARK COMPOSITION AND TAXONOMY

B.1 DATASET STATISTICS

Table 5 gives the composition of the released dataset. Its per-axis distribution is shown in Fig. 2.

Table 5: Composition of the released dataset.

Property	Main split
Total scenes	2 078
Audio–visual / audio-only	1 347 / 731
Chinese / English	1 170 / 908
Positive / no-demand negative	1 631 / 447
Generated / human-recorded interactions	1 801 / 277
Multi-turn interactions	969

B.2 POSITIVE-DEMAND SCENES: SIX AXES

Fig. 2 summarizes the taxonomy and its distribution in the released dataset. Axes 1–4 describe construction challenges; Axes 5–6 describe task and domain coverage. Categories are not an ordered difficulty scale. The Axis 1 definitions below use the audio–visual vocabulary. Audio-only scenes instead use 1.1 for language-channel demands and 1.2 for acoustic-event dependence. Their Axis 5.2 denotes acoustic perception and recognition; the remaining demand types retain their meanings.

Axis 1: Modality Dependency

Scope. This axis describes which input modalities are necessary to infer the user's intent.

1.1 Language-channel. The user's intent requires no visual or nonverbal audio grounding. It may still depend on textual dialogue history, which is classified independently on Axis 2. Typical evidence is a direct request with verbally stated arguments or a textual back-reference to an option named earlier.

1.2 Vision-grounded. The user's intent depends on visible objects, people, spatial relations, written text, gestures, or visual events in the current scene. Typical evidence includes deictic expressions such as "this one", "that person", or "the item on the left".

1.3 Audio-grounded. The user's intent depends on nonverbal sounds, environmental audio events, speaker identity, voice characteristics, or acoustic cues. Typical evidence includes asking about a sound, reacting to an alarm, or referring to someone by voice.

1.4 Vision–Audio Joint. The user's intent requires both visual and audio evidence; either modality alone would leave the intent incomplete or ambiguous. For example, "Which of these instruments makes that sound?" requires identifying the instrument from the audio and locating it among the visible candidates.

Axis 2: Contextual Disambiguation

Scope. This axis describes what kind of context is needed to resolve the user's intent beyond the literal wording of the current utterance.

2.1 Self-contained Current Context. The current utterance and scene state determine the intent without recalling earlier events or tracking changes over time.

2.2 Temporal Event Dependency. The intent requires recalling a perceptual event or tracking its timing, sequence, or resulting state change, before or during the current turn.

2.3 Dialogue History Dependency. The intent depends on prior conversational context, especially previous user–assistant exchanges.

2.4 Speaker Identity Dependency. The intent depends on identifying which person or voice is being referred to, or on distinguishing among multiple speakers.

2.5 Affective State Dependency. The intent depends on recognizing emotion, attitude, stress, fatigue, urgency, hesitation, or another affective state.

Boundary note. Axis 2.2 concerns perceptual event history, whereas Axis 2.3 concerns discourse history. A scene may contain both, but the label reflects the primary source of disambiguation.

Axis 3: Intent Expression Form

Scope. This axis describes how the user's intent is expressed linguistically.

3.1 Direct Instruction. The user states an explicit command or request for the assistant to perform an action.

3.2 Direct Question. The user expresses the intent as an explicit question.

3.3 Indirect or Descriptive Request. The user describes a situation, problem, preference, or state from which the intended assistance must be inferred.

3.4 Hesitation or Self-correction. The user's expression includes hesitation, revision, restart, or correction that must be resolved.

3.5 Interrupted or Incomplete Request. The utterance is cut off, interrupted, or incomplete, but context still determines a valid demand. R3 (App. B.3) applies when no actionable demand remains.

3.6 Compound or Multi-intent Request. The user expresses multiple related or parallel intents in one turn.

Axis 4: Acoustic Environment

Scope. This axis describes the acoustic conditions under which the user's demand is expressed. It captures speech intelligibility and audio interference, not whether audio is part of the task content.

4.1 Clean Near-field Speech. The user's speech is near-field, clear, and minimally affected by interference.

4.2 Background Noise. Non-speech environmental noise is present but does not fully obscure the user's speech.

4.3 Competing Speech. Other human speech overlaps with or competes against the user's speech.

4.4 Low-volume or Whispered Speech. The user intentionally speaks softly, whispers, or has significantly reduced volume.

4.5 Media Playback Interference. Media audio, such as television, music, phone playback, navigation audio, or public announcements, interferes with the user demand or is semantically relevant to it.

Boundary note. Axis 1 asks whether audio is part of the intent content; Axis 4 asks whether the acoustic environment affects how the intent is expressed or perceived.

Axis 5: Demand Type

Scope. This axis describes the functional type of assistance requested by the user.

5.1 Information Seeking and Retrieval. The user asks the assistant to retrieve, look up, or provide factual information.

5.2 Visual Perception and Recognition. The user asks the assistant to perceive, identify, describe, compare, or reason about visible content.

5.3 Action Execution. The user asks the assistant to perform an operation or control an external function.

5.4 Guidance and Explanation. The user asks for step-by-step guidance, procedural help, troubleshooting, or explanation.

5.5 Recommendation and Decision Support. The user asks the assistant to compare options, make a recommendation, or support a choice.

5.6 Monitoring and Reminder. The user asks the assistant to monitor a condition, detect a change, or remind them later.

5.7 Recording and Summarization. The user asks the assistant to record, transcribe, summarize, or preserve information.

5.8 Social and Emotional Interaction. The user seeks social support, emotional response, interpersonal advice, or empathetic interaction.

5.9 Translation and Language Conversion. The user asks the assistant to translate, paraphrase, rewrite, or convert language.

5.10 Content Creation. The user asks the assistant to generate new content.

Axis 6: Scenario Domain

Scope. This axis describes the everyday domain or activity setting in which the user demand occurs.

6.1 Home and Daily Life. Domestic routines and everyday household activities.

6.2 Cooking and Dining. Food preparation, eating, kitchen use, restaurants, and meal-related decisions.

6.3 Shopping and Commerce. Purchasing, product comparison, retail settings, payment, and consumer decisions.

6.4 Travel and Navigation. Trip planning, wayfinding, route selection, transportation logistics, and location search.

6.5 Tourism and Exploration. Sightseeing, visiting attractions, cultural exploration, and travel discovery.

6.6 Learning and Education. Studying, teaching, tutoring, training, and knowledge acquisition.

6.7 Office and Collaboration. Workplace tasks, meetings, documents, coordination, and professional communication.

6.8 Sports and Fitness. Exercise, training, physical activity, performance tracking, and fitness guidance.

6.9 Health and Medical Care. Health monitoring, medication, symptoms, care settings, and medical assistance.

6.10 DIY and Repair. Assembly, maintenance, repair, tool use, and practical troubleshooting.

6.11 Driving and In-vehicle Contexts. Driving, vehicle use, road context, navigation, and in-car assistance.

6.12 Creative Work and Content Production. Creative planning, media production, design, writing, and editing.

6.13 Social Gatherings. Social interaction in group settings, parties, visits, and interpersonal coordination.

6.14 Outdoor and Nature. Outdoor environments, nature activities, weather, and field observations.

6.15 Childcare and Caregiving. Caring for children, elderly people, patients, or others who need assistance.

6.16 Fashion and Beauty. Clothing, appearance, grooming, cosmetics, and style decisions.

6.17 Public Transit and Commuting. Public transportation, commuting routines, stations, transfers, and announcements.

6.18 Entertainment and Leisure. Games, media, hobbies, recreation, and leisure activities.

6.19 Safety and Emergency. Risk, danger, urgent response, accident prevention, and emergency assistance.

B.3 NO-DEMAND SCENES: SIGNAL TYPE AND INVALIDATION REASON

No-demand scenes contain cues that resemble a request, but no valid demand is directed to the conversational assistant. The two dimensions specify the demand-like signal that creates false-trigger risk and the reason it is invalid.

Dimension N1: Demand-like Signal Type

Scope. This dimension identifies the observable cue that may be mistaken for a user demand.

S1 Linguistic Signal. The utterance surface form resembles a request, question, imperative, complaint, or help-seeking expression.

S2 Semantic Signal. The utterance mentions a function, task, or topic the assistant could help with, even if the user is not actually requesting assistance.

S3 Prosodic Signal. The speech sounds as if it may be directed to the assistant because of voice quality, direction, distance, or command-like prosody.

S4 Interactive Signal. The dialogue structure resembles a user–assistant request–response turn, even though no new demand is being made.

S5 Contextual Signal. The scene presents a plausible opportunity for the assistant to help, but the user is not actually requesting assistance.

Dimension N2: Invalidation Reason

Scope. This dimension identifies why a demand-like signal is invalid.

R1 Wrong Addressee. The utterance comes from the user, but is addressed to another person, a pet, or an object, or has no intended recipient; it is not addressed to the assistant.

R2 Wrong Source. The apparent request is produced by media playback, another speaker, an echo, or another background source, rather than by the user.

R3 Incomplete or Withdrawn Demand. A possible demand is started but not completed, is withdrawn or abandoned, or is resolved by the user before it becomes actionable, leaving no actionable demand to the assistant.

R4 Non-real Intent. The utterance does not express the user's current real intent; it is hypothetical, joking, exaggerated, quoted, read aloud, rehearsed, remembered, or discussed as someone else's words.

R5 Function Mention Not Directed as a Request. The utterance discusses, evaluates, teaches, recalls, or gives feedback about an assistant capability, but does not request its use.

A negative remains demand-like in its surface cues. Labels follow the benchmark's assistant-interaction setting: the source, addressee, and current communicative intent determine whether the final turn contains a demand. In Fig. 2, Wrong Addressee and Wrong Source retain their full names; Incomplete, Non-real, and Function Mention abbreviate Incomplete or Withdrawn Demand, Non-real Intent, and Function Mention Not Directed as a Request, respectively.

C Construction and human verification

C.1 Target selection and seed sampling

Scene seeds and positive targets. Each corpus seed provides an everyday scene category and topic, such as "Office & Workplace / office plant care techniques". The generator uses this pair as a starting setting and subject, and combines it with a challenge target from Axes 1–4. For Axis 5, we supply candidate demand types; the generator chooses one that fits the developing scene when constructing the user's intent. Axis 6 is assigned from the completed scenario's domain. Thus, the seed category guides scene creation while Axis 6 describes the resulting scene, and the seed topic does not fix an Axis 5 demand type.

Coverage and seed sampling. For positive targets, the selector prioritizes uncovered admissible Axes 1–4 combinations, then those with the lowest counts, breaking ties randomly. The Axis 5 candidates include a least-used demand type and alternatives weighted toward less-used types. Seeds are sampled independently of these targets, uniformly among the least-used scene–topic pairs in the selected language. Language selection follows the 1:1 Chinese:English ratio.

Negative targets. No-demand scenes use the same scene–topic seeds with a signal–invalidation target (App. B.3), without an Axis 5 demand type. The sampler first draws an invalidation reason uniformly, then a pattern uniformly within that reason; the pattern supplies its demand-like signal type. Acoustic conditions are varied separately, and patterns that depend on a previous assistant response request dialogue history.

C.2 Script generation and review

From seed to interaction. For positives, a brainstorm combines the seed and taxonomy target into a scene concept, the concrete challenge, and an explanation of why it requires context. The cascade then specifies the physical setting, people, objects, and sound sources. Its usual path fixes the intended demand and decisive evidence, builds any required event or dialogue history around them, and finally writes the user's utterance. Patterns that require joint intent–utterance construction instead build history first and generate the two together. For negatives, a separate cascade builds a scene skeleton and any required history, then instantiates the demand-like cue and the evidence that invalidates it. In a Wrong Source scene, for example, the apparent request must remain attributed to the playback or other speaker that produced it.

Consistency and validity. Programmatic checks verify speaker roles, required history or event traces, and explicit pattern constraints. An LLM reviewer checks plausibility, causal coherence, taxonomy alignment, and support for the intended interpretation in the script. Negative validation asks whether an unmet need is present and whether it is directed to the assistant, while checking that the intended misleading signal is actually present.

Utterance-only screening. For positives, the text-only discriminator uses a single LLM call with two instructed steps: infer intent from the scripted user utterance, then compare it with the supplied intended demand. It returns a `match`, `partial`, or `mismatch` verdict, together with the inferred intent, confidence, and reasoning. Dialogue history, speaker metadata, and media evidence are withheld. A full match rejects a candidate intended to require contextual evidence, including a language-channel candidate whose difficulty should come from dialogue history. For negatives with user speech, the reverse screen uses the intent-prediction step and rejects candidates whose wording alone is judged to contain no demand; cases without a user utterance skip this check. These are script-stage screens, not tests of exclusive evidence dependence in the realized media.

C.3 Media production and annotation

Rendering and performance guides. Accepted timelines guide synthetic media production for both modalities. Multi-clip scenes are rendered sequentially with earlier clips as video references for consistent people, voices, and surroundings. Audio-only production retains the audio track. Performance guides for human-recorded interactions specify roles, dialogue, event order, and essential cues, allowing practical setting or prop substitutions that preserve the challenge. Single-turn wording may vary while retaining the intended ambiguity; multi-turn guides preserve user lines that match prepared assistant replies. Both routes enter media-grounded annotation to account for departures from the script.

Release media format. For release, audio–visual clips are encoded as H.264 video with AAC audio at 48 kHz, with the shorter spatial dimension capped at 720 pixels while retaining the aspect ratio. Audio-only clips are distributed as WAV files without this video-transcoding step. Frame rates are not standardized to a single value across the release.

Media-derived draft and script reconciliation. An initial annotation uses multimodal media captions and word-level ASR, without the script's intended annotation. Captions supply scene and speaker semantics; ASR supplies spoken wording and timing. The next stage checks scripted speech against ASR and compares the draft with the scenario specification. Significant disagreements become observable questions for targeted MLLM inspection of the media. The draft, actual script, ASR, and resulting observations are then reconciled. In multi-clip scenes, earlier user clips and assistant replies provide history; annotation and timing target the final user clip.

Timing, relevance, and verification. Segment transcripts are matched to ASR words, and their start and end times are aligned to word boundaries. Rule-based and LLM checks remove context that does not materially affect the appropriate response, while retaining referents needed to interpret the demand. Schema and cross-field checks, together with an LLM review, assess demand presence, temporal grounding, intent, context quality, and transcript coverage before key points are constructed.

Key-point construction and grounding. Verified intent and context are decomposed into atomic, response-relevant key points, with the transcript helping identify their evidence source. Each point must be expressible from the structured intent or an existing required-context description. Points representing the requested task, required context, or both receive the corresponding labels described in App. A. An MLLM then checks the actual media against the points, context, and intent. Unsupported details are removed or narrowed to supported content. If this changes the intent, key points are regenerated and checked against the media again.

C.4 HUMAN VERIFICATION

Media quality. Human reviewers are asked to assess perceptual realism and playback usability of the generated media. Generated clips were rejected for conspicuous synthesis artifacts, such as mismatched lip movements, sudden object appearances or disappearances, and distorted faces or physically implausible motion. Across both media sources, rejection criteria also included corrupted video or audio, missing sound, and severe playback stalls. Unintended assistant-response audio and crosstalk that made speakers indistinguishable were also grounds for rejection.

Annotation correctness. Reviewers checked each applicable machine-produced annotation field against the realized interaction: demand presence, demand spans, transcript, structured intent, required context, user profile, and key points. They verified speech and timing against the media and checked semantic annotations for unsupported details, irrelevant context, and missing key points, using dialogue history where needed. Generated scenes were retained only when all applicable fields passed review. For human-recorded interactions, experts corrected erroneous fields before finalization to accommodate departures from the performance guide.

Figure 7 illustrates the review controls using a generated benchmark example in place of recorded media. Two consecutive views show the complete annotation form in its original field order, with the reference annotation expanded. The example uses the released English annotation; selected controls show interface defaults rather than a saved human verdict.

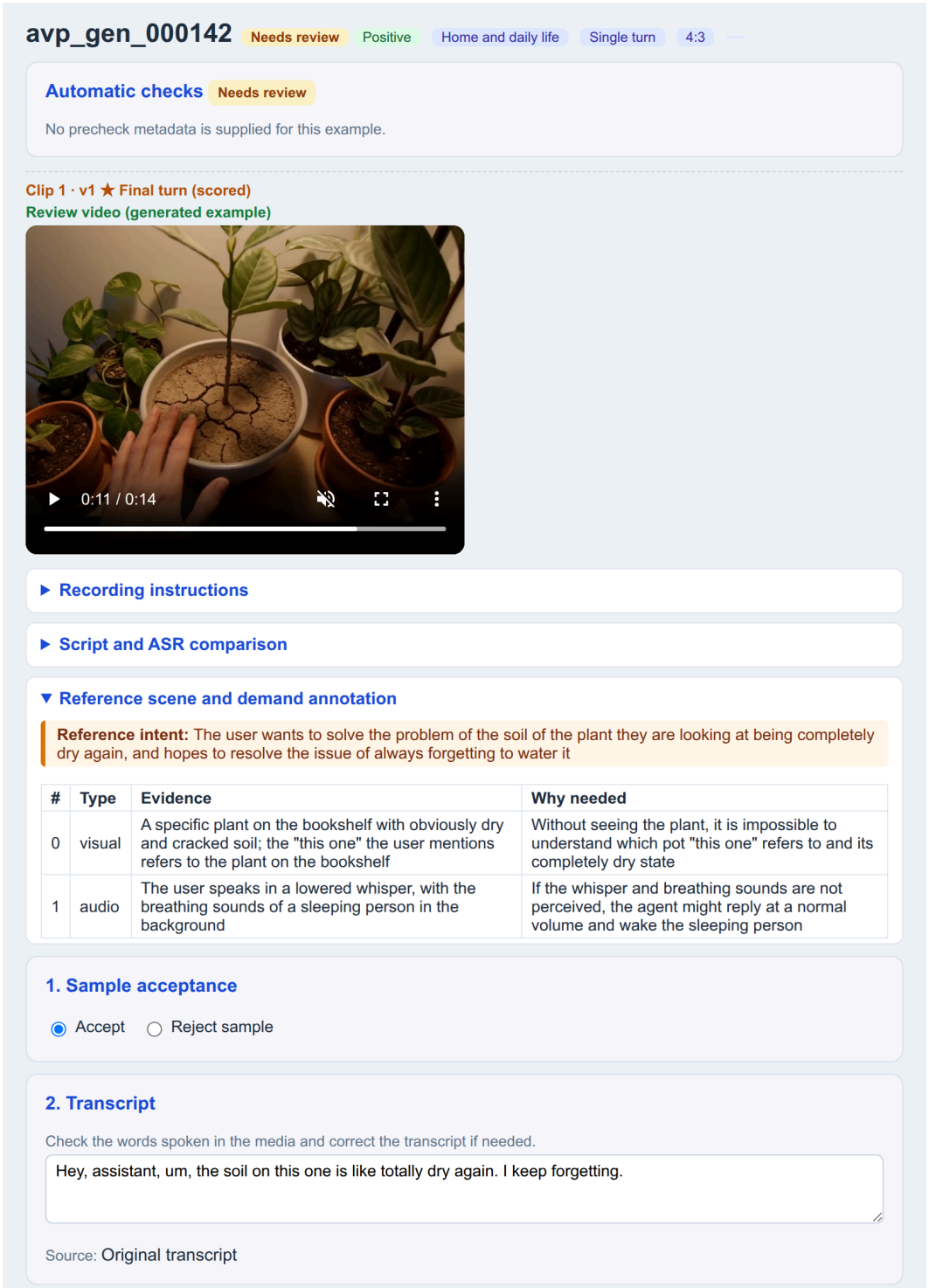


Figure 7: Annotation-review interface, shown in two consecutive views. (a) Generated media, reference annotation, sample acceptance, and transcript. The original field order is retained; selected options are interface defaults.

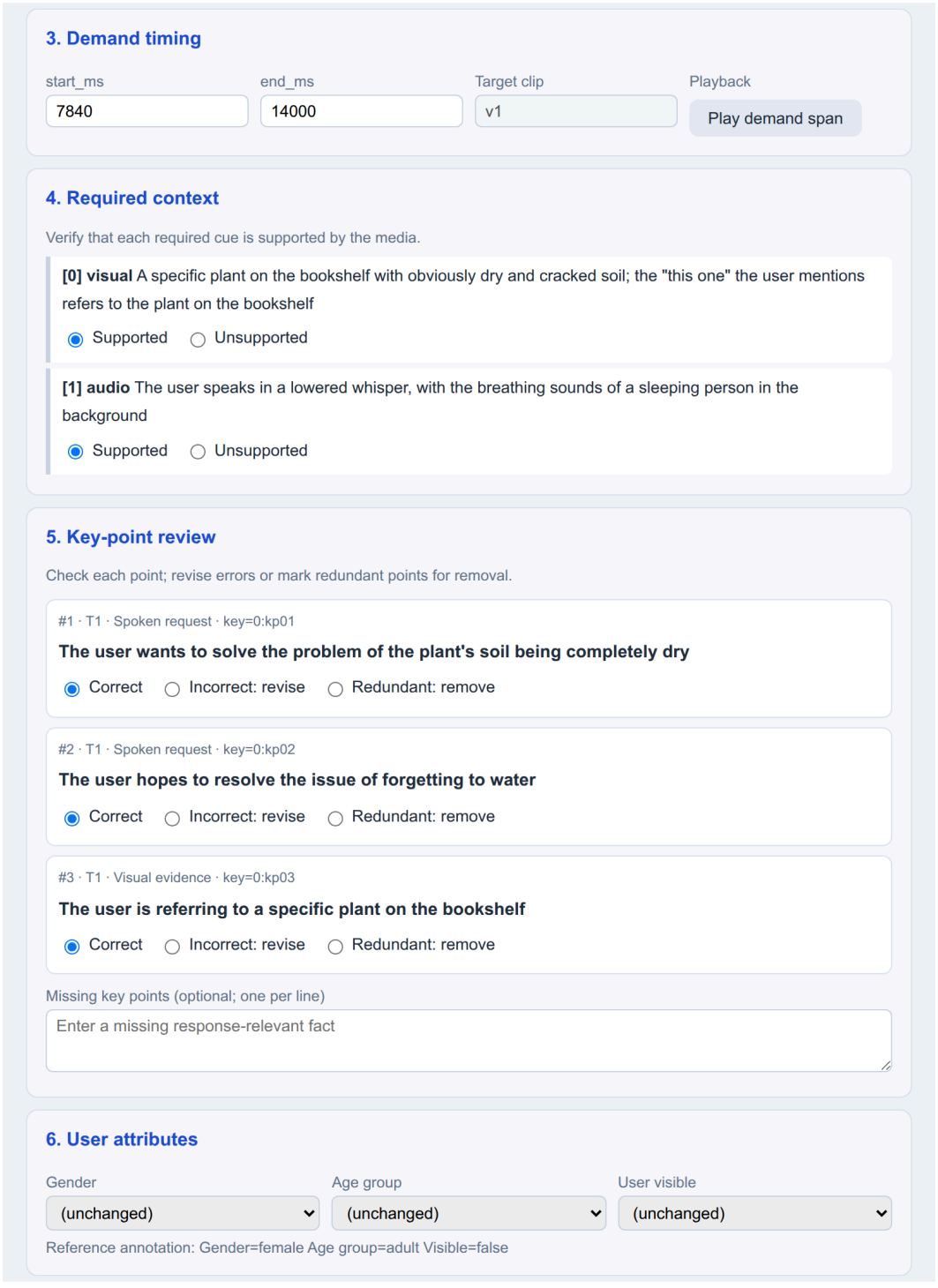


Figure 7: Annotation-review interface (continued). (b) Demand time spans, required context, all key points, and user profile fields from the same generated sample.

D EXPERIMENTAL SETUP

D.1 INPUT PREPARATION AND EVALUATION

User-turn media are presented in chronological order, with preceding assistant replies supplied as text between turns. Models predict the demand in the final user clip; preceding clips and assistant replies provide interaction history. Predicted start and end times are local to the final clip. The model-facing instruction is specialized to audio–visual or audio-only input, and predictions use the same demand-decision schema and scoring procedure. The full prediction instruction is provided in App. F.1.

Media-source comparison. The comparison in §5.5 and Table 4 uses only Chinese audio–visual scenes from each source, restricting both the synthetic and human-recorded subsets to the same language and input modality.

D.2 TEXT-ONLY BASELINE

The text-only baseline uses GPT-5.4 with the same modality-specific task instruction (App. F.1) as the native systems. Each media block is replaced in place by its ASR transcript, sentence-level timestamps, and optional speaker identifiers. Multi-turn inputs retain the original clip order, turn labels, and preceding assistant reply text. Each replacement block explicitly states that native media are unavailable. The baseline is evaluated with the same parser, segment matching, and metrics as the native systems.

System message: The full prediction instruction in App. F.1, including the demand annotation schema and field constraints; the audio-only variant follows App. F.2.

User message: The panels below show single-turn and multi-turn input templates, followed by input examples for a positive and a no-demand scene. Demand labels reflect the full scenes and may not be recoverable from the ASR transcripts alone. The templates retain the original audio–visual task wording, including "video", while replacing media with ASR text. For audio-only input, "video" and "audio or video" are replaced with "audio".

Braces mark placeholders. The `duration_note` adds " (clip duration {duration_ms} ms)", and `speaker_id` denotes the speaker identifier returned by ASR.

The multi-turn template shows the complete ASR block for each clip, including its no-media notice. ASR blocks and assistant replies appear in chronological order; longer histories repeat the same pattern. If no speech is recognized, the no-speech notice replaces that clip's ASR block. The closing instruction follows the final block in either case.

User-input templates

Single clip

[No audio or video is provided for this clip. Below is the verbatim ASR transcript of its speech, with millisecond timestamps local to this clip{duration_note}.]

[{begin_ms}-{end_ms} ms] speaker {speaker_id}: {ASR sentence}

...

Annotate the user's demand in this video and output the demand-decision JSON.

Replacement ASR block when no speech is recognized

[No audio or video is provided for this clip, and automatic speech recognition returned no speech for it.]

User-input template with dialogue history

You will receive multiple video clips from the same scene in chronological order. Together they form the dialogue history; the last video clip is the current user turn to annotate:

User-input template with dialogue history (continued)

[Turn 1 · user video]

[No audio or video is provided for this clip. Below is the verbatim ASR transcript of its speech, with millisecond timestamps local to this clip{duration_note}.]

[{begin_ms}-{end_ms} ms] speaker {speaker_id}: {ASR sentence}

...

[Turn 1 · agent reply] {preceding assistant reply}

[Turn 2 · user video (final turn, annotate this)]

[No audio or video is provided for this clip. Below is the verbatim ASR transcript of its speech, with millisecond timestamps local to this clip{duration_note}.]

[{begin_ms}-{end_ms} ms] speaker {speaker_id}: {ASR sentence}

...

Annotate only the user's demand in the final video clip and output the demand-decision JSON. Use the preceding video clips and agent replies only as dialogue-history context; do not output segments for them. start_ms/end_ms must be local timestamps within the final video clip.

Positive example (Demand present)

[No audio or video is provided for this clip. Below is the verbatim ASR transcript of its speech, with millisecond timestamps local to this clip (clip duration 9055 ms).]

[5040-8640 ms] speaker 0: Set it up, so when he gives the verbal, okay, export this version.

Annotate the user's demand in this video and output the demand-decision JSON.

Negative example (No demand)

[No audio or video is provided for this clip. Below is the verbatim ASR transcript of its speech, with millisecond timestamps local to this clip (clip duration 12051 ms).]

[4590-10710 ms] speaker 0: my data was leaked, can this it insurance immediately activate an emergency response for me?

[11070-11990 ms] speaker 0: how do i file a claim?

Annotate the user's demand in this video and output the demand-decision JSON.

D.3 Semantic scoring

We use Qwen3.6-Flash with thinking enabled as the semantic judge in the main experiments. The judge receives the predicted structured intent concatenated with the predicted required-context descriptions, together with the reference key points. Reference transcripts and context-necessity notes are not added to the candidate answer text. App. F.4 provides the judge instruction, and App. E.2 analyzes result stability across judges.

Each distinct reference point is counted once, regardless of tier labels. Key points in unmatched reference segments count as missed. Segment matching and score aggregation follow App. A.

E Supplementary experiments and diagnostics

This appendix provides supplementary analyses of request and context coverage, LLM judge stability, and positive-demand challenges. It then presents the blinded comparison of responses generated with and without demand annotations and closes with qualitative error cases.

E.1 Request and context key-point coverage

We compare recovery of spoken requests and the context needed to interpret them, extending the analysis in §5.3 and Figure 4.

Population and counting. We analyze 1 060 audio–visual demand scenes scored by all four systems, containing 3 969 distinct reference key points. Each point counts once; points in unmatched reference segments count as misses. Context comprises visual, audio, dialogue-history, and other context points (emotion, activity, identity, and scene condition). All context pools these categories.

Coverage gaps and sampling uncertainty. Table 6 uses the same population and point weighting as Figure 4. Coverage is the ratio of recovered to reference key points in each channel. The gap subtracts context coverage from spoken-request coverage, in percentage points. Confidence intervals assess sensitivity to the composition of the sampled scenes.

Scene-bootstrap procedure. For each of 10 000 resamples, we draw 1 060 scenes uniformly with replacement from the common population. All systems and channels use the same draw, and each selected scene contributes all its key points, including repeated contributions when selected more than once. Model outputs and judge verdicts remain fixed.

Within each resample, we recompute each channel's coverage from its pooled hit and reference-point counts, then calculate the request-minus-context gaps. Their 2.5th and 97.5th percentiles form the 95% confidence interval. The Gap column uses the original full sample; resampling supplies the interval while preserving dependence among each scene's key points.

Table 6: Point-weighted coverage by evidence channel (Figure 4). Gap subtracts each context row's coverage from the same model's Spoken request coverage; positive values mean lower context coverage. The 95% confidence interval quantifies the gap's uncertainty under scene resampling. Both are in percentage points (pp). n counts reference points; dashes mark the request reference. Gaps use unrounded rates.

System	Channel	Reference points (n)	Coverage (%)	Gap (pp) Request − context	95% confidence interval for gap (pp)
Gemini 3.1 Pro	Spoken request	1 984	82.9	–	–
	All context	1 972	44.7	38.2	[35.4, 41.1]
	Visual	804	48.6	34.2	[30.3, 38.3]
	Audio	557	47.2	35.6	[31.2, 40.2]
	History	475	42.1	40.8	[35.9, 45.7]
Qwen3.5-Omni-Plus	Spoken request	1 984	83.6	–	–
	All context	1 972	48.9	34.6	[32.0, 37.3]
	Visual	804	47.3	36.3	[32.6, 40.1]
	Audio	557	45.6	38.0	[33.6, 42.3]
	History	475	57.7	25.9	[21.1, 30.6]
Seed 2.0 Lite	Spoken request	1 984	76.0	–	–
	All context	1 972	52.8	23.2	[20.4, 26.0]
	Visual	804	54.7	21.3	[17.4, 25.1]
	Audio	557	46.5	29.5	[25.2, 34.0]
	History	475	62.3	13.7	[8.8, 18.5]
GPT-5.4 + transcript	Spoken request	1 984	78.9	–	–
	All context	1 972	33.0	45.9	[43.2, 48.6]
	Visual	804	20.3	58.6	[55.2, 61.9]
	Audio	557	33.6	45.3	[41.3, 49.4]
	History	475	58.7	20.1	[15.4, 24.9]

Results and interpretation. Across the four systems, request coverage exceeds pooled context coverage by 23.2–45.9 percentage points. Every reported context-channel gap has a confidence interval entirely above zero, supporting lower context coverage after accounting for the estimated scene-sampling variation.

Audio-only extension. The same direction holds for Gemini 3.1 Pro on the 570 common audio-only demand scenes: it recovers 83.5% of request points and 49.9% of audio points. Across all 1 631 demand-bearing scenes in ODU-Bench, Gemini 3.1 Pro achieves 45.6% point-weighted coverage of contextual key points.

Textual clues and inference in the text-only baseline. We examine whether the text-only baseline's visual coverage is associated with lexical clues in its input, extending §5.3. This input includes preceding user transcripts and assistant replies (App. D.2).

Lexical overlap is the fraction of a reference point's distinct content words present in the actual ASR, dialogue, and template text. To compare with English reference points, we restrict to Latin-script inputs, yielding 345 visual points. Using saved outputs and judge verdicts, we compare coverage for points with at least 20% overlap and those below this threshold.

Text-only coverage is 32.3% in the higher-overlap group and 10.7% in the lower-overlap group. Seed 2.0 Lite, which receives audio and video, scores 53.8% and 52.9%, respectively. The contrast between overlap groups is much larger for the text-only baseline. The text-only system's recovery of multimodal context may stem from textual clues in transcripts and dialogue history, combined with commonsense knowledge.

E.2 LLM JUDGE STABILITY

We examine whether reference key points can be recovered from their source annotation text and whether the ordering of three evaluated systems is stable across three LLM judges.

Design. We re-judge a fixed population with Qwen3.6-Flash, DeepSeek-V4-Flash, and GPT-5.4. The population contains 1 060 audio–visual demand scenes scored by all three evaluated systems. For each setting, all judges receive the same scoring instruction, candidate text, and reference points. Rates are point-weighted and use the all-judge common verdict set for each evaluated system; they are not the scene-averaged M2 values in Table 3.

Table 7: LLM judge stability. Panel A tests whether benchmark key points can be recovered from their source annotations. Panel B re-judges identical stored outputs; all three judges yield the same model ordering.

Setting	Quantity	Judge			Stability readout
		Qwen3.6	DeepSeek	GPT-5.4	
	Panel A: target recoverability				
A	Recoverability	0.998	0.996	0.996	3 956 key points
A	Orphan rate	0.002	0.004	0.004	ceiling loss
	Panel B: stored-output re-judging				
B	Gemini 3.1 Pro coverage	0.648	0.639	0.609	range 0.039
B	Qwen3.5-Omni-Plus coverage	0.682	0.668	0.641	range 0.041
B	Seed 2.0 Lite coverage	0.667	0.656	0.638	range 0.029
B	Model ordering	identical for 3 of 3 judges			same conclusion

Note. Rates are point-weighted proportions on the 0–1 scale. Orphan rate is one minus recoverability; range is the maximum minus minimum coverage across judges.

Annotation-to-key-point recoverability. We present each judge with the reference intent and required-context text from which the key points were derived, and measure recovery of those points. Across 3 956 key points, Qwen3.6-Flash, DeepSeek-V4-Flash, and GPT-5.4 achieve recoverability rates of 0.998, 0.996, and 0.996, respectively; their orphan rates are only 0.002, 0.004, and 0.004. This establishes high recoverability from the source annotation across the tested judges.

Model-ordering stability. We next re-score the exact stored per-segment outputs of Gemini 3.1 Pro, Qwen3.5-Omni-Plus, and Seed 2.0 Lite. No model inference or segment matching is repeated. The audit retains points with tier labels and available candidate text. Although judges differ slightly in absolute calibration, all three produce the same model ordering, Qwen3.5-Omni-Plus $>$ Seed 2.0 Lite $>$ Gemini 3.1 Pro, on the retained point sets.

E.3 POSITIVE-DEMAND CHALLENGE BREAKDOWN

We examine which types of valid demand are associated with lower M2 key-point coverage and whether these patterns recur across systems and modalities. This comparison groups whole scenes by challenge category, complementing the key-point-level analysis in App. E.1.

Design. We reanalyze saved scores from 13 systems on audio–visual (AV) scenes and 15 systems on audio-only (AO) scenes. Within each modality, we retain the same positive scenes with complete scores for all displayed systems. The four challenge axes in App. B.2 are modality dependency, contextual disambiguation, intent expression form, and acoustic environment. The modalities are analyzed separately because Axis 1 uses modality-specific categories.

Scoring follows Table 3, with missed reference points counted as misses. Each cell in Fig. 8 and Fig. 9 subtracts the system's mean M2 over all common scenes from its mean within a category, weighting scenes equally. Differences are in percentage points: red means below the system's own average and blue means above it. The *Panel mean* row averages the differences equally across all displayed systems, including the text-only comparator, which is shown separately and remains unranked.

Intent expression. Indirect or descriptive requests (Axis 3.3) fall below every system's mean in AV and AO, with panel-average M2 differences of -7.2 points (AV) and -8.2 points (AO). Interrupted or incomplete requests (Axis 3.5) also have negative panel-average M2 differences (-3.2 AV; -3.9 AO). Compound or multi-intent requests (Axis 3.6) have positive panel-average M2 differences of $+4.0$ and $+4.0$ points, respectively; the difference is positive for every AO system but varies in sign across AV systems. The shared indirect-request deficit and negative mean differences for incomplete requests point to recovering unstated intent as a recurring difficulty. Positive mean differences for multi-intent requests show that several demands need not imply lower coverage; taxonomy categories do not define an ordered difficulty scale.

Contextual disambiguation. On AO scenes, speaker-identity and affective-state dependencies have panel-average M2 differences of -7.3 and -6.1 points, respectively. For scenes that depend on dialogue history, the difference between category M2 and each system's overall mean averages $+2.0$ percentage points in AV and $+2.4$ in AO. The higher scores may stem from explicit clues in earlier turns, such as the user's stated goal or previously discussed options, that help clarify the current request.

Acoustic conditions. Acoustic categories show less consistent patterns across modalities. Background-noise scenes, for example, have a panel-average M2 difference of $+1.9$ points in AV and -0.8 points in AO. The indirect-request deficit is thus more consistent across the tested systems and modalities than the acoustic patterns.

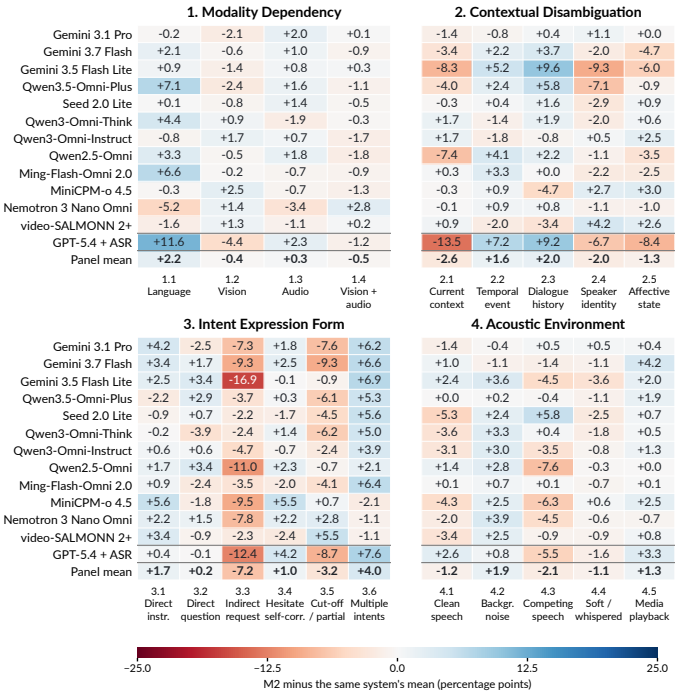


Figure 8: **Positive challenge profiles: audio–visual.** Category-minus-system-mean M2 on the common positive cohort. Labels give category codes and names; the final row is the unweighted mean across the displayed systems. The color scale is shared with Figure 9.

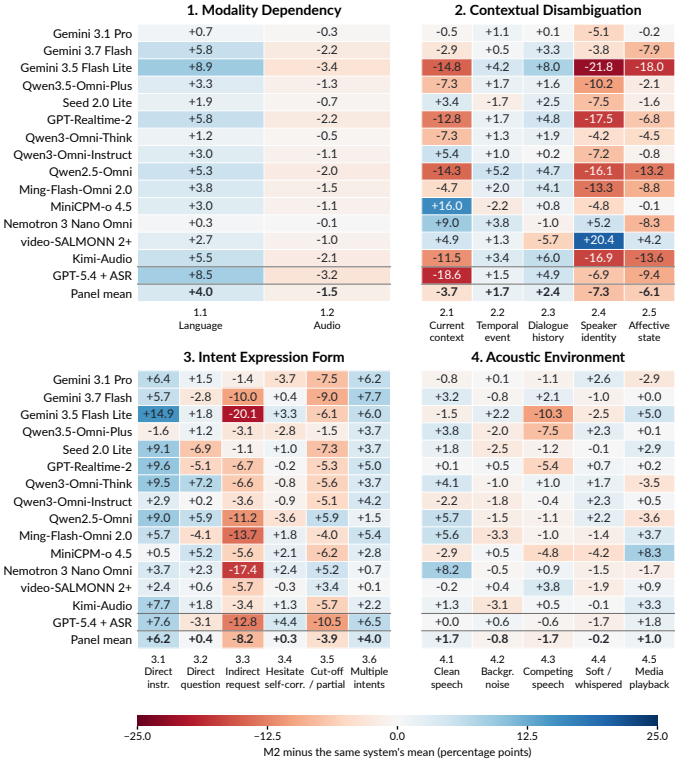


Figure 9: **Positive challenge profiles: audio-only.** Cells show category mean M2 minus each system's mean on the common positive cohort, in percentage points. The panel includes GPT-Realtime-2 and Kimi-Audio-7B-Instruct. The final row averages the displayed systems equally.

E.4 BLINDED RESPONSE COMPARISON

We assess whether providing demand annotations improves response content and the decision to respond.

Scene sampling. We sampled 200 audio–visual scenes from the human-verified benchmark. The sample comprises 160 demand scenes and 40 no-demand scenes, with Chinese and English equally represented within each group.

Paired response generation. Gemini 3.1 Pro generated a reply under each of two conditions: "Not given demand annotation" (*Not given*) and "Given demand annotation" (*Given*). Both received identical video and audio, in clip order with intervening assistant replies, and responded to the final user turn. The *Given* condition additionally received the reference structured intent and required-context descriptions joined by `|`, or "Demand: NONE" for no-demand scenes; neither key points nor reference answers were supplied. Settings were fixed: temperature 0, top-p 1, an output-token limit of 8 192, and 1 candidate per request.

Blinded assessment. The interface (Fig. 10) shows each interaction with replies A and B. Scene order and reply placement were randomized. Human assessment considered correctness, relevance, contextual grounding, and appropriate silence. Ties covered equally good or poor replies; instructions discouraged preference based on length alone.

Preference results. Table 8 reports all 200 judgments. Of these, 80 (40.0%) were ties. Among the remaining 120 comparisons, 103 (85.8%) favored *Given*. The preference advantage appears in both languages and in both demand-presence groups.

Table 8: Human response preferences: counts (percentages) within each subset. "Given" and "Not given" abbreviate the "Given demand annotation" and "Not given demand annotation" conditions. Ties remain in the denominator.

Subset	Scenes	Given preferred	Tie	Not given preferred
All	200	103 (51.5%)	80 (40.0%)	17 (8.5%)
Demand present	160	85 (53.1%)	58 (36.3%)	17 (10.6%)
No-demand	40	18 (45.0%)	22 (55.0%)	0 (0.0%)
Chinese	100	47 (47.0%)	45 (45.0%)	8 (8.0%)
English	100	56 (56.0%)	35 (35.0%)	9 (9.0%)

Response content and whether to respond. Among 148 demand scenes where both conditions produced text, *Given* was preferred in 74 cases versus 17 for *Not given*, with 57 ties. On no-demand scenes, the proportion of textual replies fell from 60.0% to 15.0%. On demand scenes, silent replies fell from 12 to 1. These findings support the utility of explicit demand and context for response content and the decision to respond.

Input prompts. The panels give the shared instruction, templates, and complete examples. Video placeholders denote media attachments; an exact `[SILENCE]` output is displayed as no reply.

Shared system instruction

Reply directly to the user's request in the last clip, using the user's language. If a demand annotation is provided, follow it; NONE means no request. If there is no request for you, output only [SILENCE].

User-input templates

Not given demand annotation	Given demand annotation
Clip 1	Clip 1
[video input]	*[video input]*
	Demand: {*structured_intent*} | {*required_context.description*} | ...

Positive example (Demand present)

Not given demand annotation	Given demand annotation
Clip 1	Clip 1
[video input]	*[video input]*
	Demand: The user wants the device that is continuously emitting a beeping sound to stop its alarm. | The continuous electronic beeping sound in the background. | The device with a blinking red indicator lamp on the mantel, which is the source of the alarm sound referred to by the user's "That beeping". | The user's tone is fatigued and irritable, and he startles when the vase shatters.

Negative example (No-demand)

Not given demand annotation	Given demand annotation
Clip 1	Clip 1
[video input]	*[video input]*
	Demand: NONE

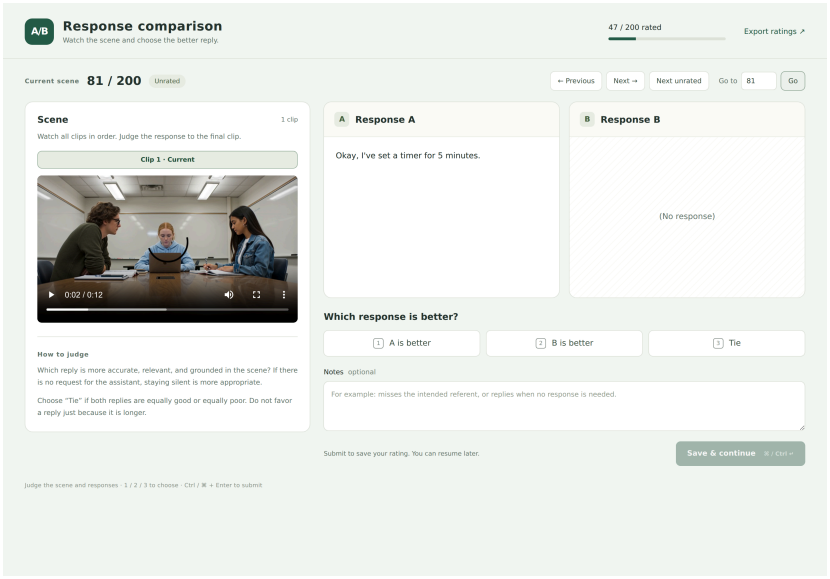
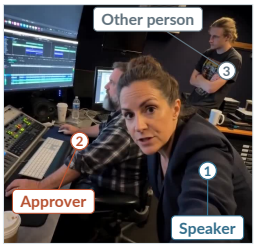


Figure 10: Blinded response-comparison interface. The human reviewer is asked to choose a preferred reply between A and B.

E.5 QUALITATIVE ERROR ANALYSIS

(a) Missing approver — DEMAND PRESENT

SCENE AND REFERENCE ROLES



TRANSCRIPT

Set it up so when he gives the verbal okay, export this version.

REFERENCE KEY POINTS

- KP1 The user wants to set up a conditional export task
- KP2 The export trigger condition is someone's verbal approval
- KP3 **The person giving approval is the male editor at the main control console**
- KP4 What is exported is the video editing version on the monitor

MODEL PREDICTIONS — Full intent + context

Gemini 3.1 Pro Demand: yes M2: 3/4 KP1–4 ✓ ✓ ✗ ✓	The user wants the assistant to prepare to export the currently selected version as soon as a verbal confirmation is given. *Visual: User is in an editing suite with project files*
Seed 2.0 Lite Demand: yes M2: 3/4 KP1–4 ✓ ✓ ✗ ✓	The user wants the system configured to export the current editing version automatically upon receiving the verbal confirmation. *Visual: Scene is a post-production editing suite with editing timeline displayed on monitors*
Qwen3.5-Omni-Plus Demand: yes M2: 3/4 KP1–4 ✓ ✓ ✗ ✓	The user wants to configure the software to automatically export a specific version upon receiving a verbal confirmation. *Visual: User points at video editing software on screen* *Activity: User is overseeing a video editing process*

ERROR ANALYSIS — **The models preserve the export condition but omit who must approve it.**

(b) Missing naming rule — DEMAND PRESENT

SCENE AND AUDIO SOURCE



Prior assistant reply
Protocol set: 'Ambient Draft' + timestamp.

TRANSCRIPT

That ambient texture, tag it.

REFERENCE KEY POINTS

- KP1 The user wants to tag and save the ambient audio clip.
- KP2 The user is referring to the currently playing ambient music.
- KP3 **Named as Ambient Draft plus a timestamp.**

MODEL PREDICTIONS — Full intent + context

Gemini 3.1 Pro Demand: yes M2: 2/3 KP1–3 ✓ ✓ ✗	The user wants the AI to tag the ambient texture audio segment currently playing. *History: Refers to agreed upon audio tagging protocol*
Seed 2.0 Lite Demand: yes M2: 0/3 KP1–3 ✗ ✗ ✗	The user wants to inquire about the component/port he is pointing to on the desktop audio device *Visual: The user points to a port on the audio device under the two desktop speakers* *History: Prior discussion about setting up naming convention for saving audio clips*
Qwen3.5-Omni-Plus Demand: yes M2: 3/3 KP1–3 ✓ ✓ ✓	The user wants to save the currently playing audio segment as an 'Ambient Draft' with a timestamp. *History: Prior agreement on naming convention and trigger phrase.* *Audio: Ambient sound currently playing through speakers.*

ERROR ANALYSIS — **The task is recovered without the agreed naming rule.**

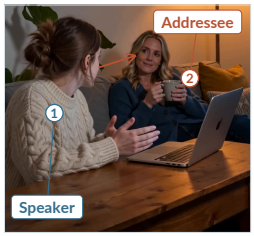
Gemini 3.1 Pro recovers the audio-tagging task but omits the agreed name and timestamp. Qwen3.5-Omni-Plus preserves both. Seed 2.0 Lite instead describes a question about a device port, changing the task itself.

Figure 11: Qualitative error analysis on English generated audio–visual scenes. (a)–(b) Demand-present cases; M2 reports covered/reference key points.

(c) Wrong addressee

NO DEMAND

SCENE AND REFERENCE ROLES



TRANSCRIPT

Can you look up what a typical day for a data analyst looks like? I need an encyclopedia-styled breakdown of the tasks and meetings.

REFERENCE DECISION

The request is addressed to the companion, not the evaluated assistant.

The speaker's body orientation, eye contact and conversational gestures point to her companion.

MODEL PREDICTIONS — Full intent; context empty in all outputs

Gemini 3.1 Pro	Demand: yes M2: N/A False trigger	The user wants an encyclopedia-style breakdown of a typical day for a data analyst, including tasks and meetings.
Seed 2.0 Lite	Demand: yes M2: N/A False trigger	The user wants to find information about what a typical day for a data analyst looks like, and get an encyclopedia-style breakdown of their tasks and meetings.
Qwen3.5-Omni-Plus	Demand: yes M2: N/A False trigger	The user wants to get an encyclopedia-styled breakdown of the daily tasks and meetings of a data analyst.

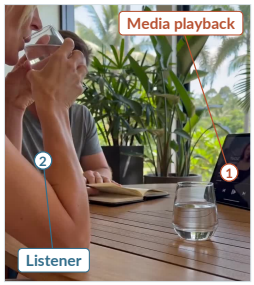
ERROR ANALYSIS — **The intended addressee is lost.**

All three models turn the spoken topic into an assistant task. The speaker's orientation and eye contact show that she is addressing her companion; the outputs fail to preserve this distinction.

(d) Playback false trigger

NO DEMAND

MEDIA PLAYBACK AND LISTENERS



TRANSCRIPT

how can i control my carb intake without feeling hungry all the time? Could you help me check if this recipe is reasonable? Please recommend some good snacks.

REFERENCE DECISION

The requests come from the tablet's media playback.

The people present remain listeners.

MODEL PREDICTIONS — Full intent + context

Gemini 3.1 Pro	Demand: yes M2: N/A False trigger	Segment 1 The user wants advice on controlling carb intake while avoiding hunger and asks the assistant to evaluate a specific recipe. *Visual: User refers to a recipe shown in the notebook.* Segment 2 The user wants the assistant to recommend some good snacks. *History: User previously asked about carb intake.*
Seed 2.0 Lite	Demand: yes M2: N/A False trigger	The user wants to know how to control carb intake without constant hunger, get a check on the reasonableness of the recipe, and receive recommendations for good snacks. *Visual: The man beside the user is looking at the notebook that contains the recipe to check*
Qwen3.5-Omni-Plus	Demand: yes M2: N/A False trigger	The user wants snack recommendations.

ERROR ANALYSIS — **Playback is promoted into a user demand.**

The tablet plays diet questions while the woman drinks water. All three models treat this playback as a user request; Gemini 3.1 Pro and Seed 2.0 Lite also link the recipe question to the physical notebook.

Figure 11: Qualitative error analysis (continued). (c)–(d) No-demand cases.

F EVALUATION PROMPTS

This section presents the prediction and key-point judging instructions distributed with the evaluation code. Scenario-generation and annotation-construction prompts, including key-point extraction and grounding, are not distributed; their procedures are described in App. C. The audio–visual prediction instruction is reproduced in full, followed by the audio-only differences and a pointer to the text-only input templates. Executable AV and AO variants are provided with the evaluation code.

F.1 SHARED PREDICTION TASK

In the reproduced instruction, "the last segment" refers to the final input clip; output segments are demand-bearing regions within that clip.

Audio-Visual Prediction Instruction

PURPOSE	The prediction instruction for an audio-visual scene. It asks the model to decide whether the final user turn carries a demand and, when it does, to localize the demand in time, transcribe it, state a self-contained intent, and list the multimodal context the demand depends on.
INPUTS	One video clip per user turn, in order, with the assistant replies that separate them supplied as text.
OUTPUT	One JSON object with `has_demand` and a `segments` list. Each segment carries `start_ms`, `end_ms`, `transcript`, `structured_intent`, `required_context` and the user attribute fields.
VARIANT	Complete prompt; audio-visual variant.

You are the user's AI assistant. You will watch one or more videos, in which the user may make requests to you.

Task

Determine whether a user has made a request to you in the video. The scenario may be a user interacting with you or ordinary interpersonal conversation – you must distinguish between them based on the visual and audio content. If the user expresses a demand, output the following information in structured JSON format (instead of replying directly to the user). A demand is not necessarily an explicit question or instruction; it may also be an implicit expression (e.g., complaining, talking to oneself, hinting for help, etc.).

- **Time Localization**: The start and end times when the demand occurs (milliseconds)
- **`transcript`**: The exact words spoken by the user within this segment's time interval (verbatim transcription, no additions or omissions; does not include utterances from other speakers, nor content from previous or subsequent turns)
- **Structured Intent**: A single sentence summarizing the user's goal, starting with "The user wants to...", semantically complete and self-contained
- **Required Context**: The multimodal context information necessary to understand the content of this demand
- **Is the user in the frame?**
- **User gender and age group**

Required Context (`required_context`)

Fill in context items where "the absence would lead to an incomplete understanding of the demand, a deviation in the response direction, or a significant decrease in response quality". If the user's utterance text alone is sufficient to accurately and appropriately understand and respond, then `required_context = []`. Select `type` from the following 7 options:

Audio-Visual Prediction Instruction (continued)

type	Meaning
`visual`	Objects/text/actions/scenes in the visual frame are necessary information for understanding the demand
`audio`	Ambient sounds/background audio provide clues for understanding the object or context of the demand
`dialogue_history`	The current utterance contains references or omissions, requiring previous dialogue content for comprehension
`user_activity`	The user's ongoing activity provides necessary context for understanding the demand
`user_identity`	The user's identity/role/professional background affects the understanding of the demand
`user_emotion`	The user's emotion/tone affects the judgment of intent direction (e.g., a dissatisfied tone turns an observation into a request)
`scene_condition`	Environmental conditions (time/weather/noise, etc.) affect the understanding of the demand

Fill in a concise `description` for each entry, retaining all details needed to identify the referent and understand the demand. Each type must appear at most once.

Multi-turn Input

The input may contain multiple video clips (`v1`, `v2`, `v3`...), each corresponding to a user turn, with possible text replies from the AI agent between clips.

- **Annotate user demands only in the last segment**. Content from previous turns and agent replies serve only as dialogue history context; do not output them as segments.
- `has_demand` is `true` only when a user demand exists in the last segment.
- `start_ms`/`end_ms` are relative to the start of the last segment, not global time.

Segment Division

One segment corresponds to a continuous region of demand expression. Boundaries are determined solely by the following two signals:

- **Must split** (highest priority): Within the final user clip, split at speech interruptions: at least 2 seconds of silence or an insertion of speech unrelated to the demand, such as talking to bystanders.
- **Do not split** (only within the same turn and the same continuous utterance): Consecutive multiple intents, self-correction, preamble + request, emotion + action, etc., must all be kept as a single segment.
- Express `structured_intent` for compound demands using a compound sentence (e.g., "The user wants to know the weather and set an alarm"); for self-correction, record only the final valid demand.
- **transcript boundaries**: `transcript` must strictly contain only the speech transcription of *the user* within the `start_ms` to `end_ms` time window. Do not include utterances from other speakers, AI replies, or content from previous or subsequent turns in the transcript. If preceding information is relevant to demand comprehension, place it in `required_context` (type=`dialogue_history`) instead of the transcript.

Field Constraints

The following fields are closed-set classifications. Use exactly one of the given values; do not invent synonyms (for example, do not write `"young adult"`, `"senior"`, `"middle-aged"` or `"teenager"`).

- `user_in_frame`: boolean. Whether the user initiating the demand is within the frame.
  - `true`: the user initiating the demand is visible in the scene.
  - `false`: first-person off-screen voice, only the user's voice is heard, or it is impossible to determine who in the scene initiated the demand.
- `user_gender`: string, must be one of the following values:

Audio-Visual Prediction Instruction (continued)

  - `"male"`: there is clear evidence of male or masculine presentation.
  - `"female"`: there is clear evidence of female or feminine presentation.
  - `"unknown"`: evidence is insufficient, evidence from multiple people/speakers is mixed together or voice/visual evidence cannot support a judgment, or no inference should be made.
- `user_age_group`: string, must be one of the following values:
  - `"child"`: child or clearly underage person.
  - `"adult"`: adult.
  - `"elderly"`: elderly person, with clear age features.
  - `"unknown"`: evidence is insufficient or judgment is impossible.

Output Format

Output only JSON, with no comments or code block markers. The timestamp values in the examples below are placeholders only; infer the real `start_ms`/`end_ms` from the provided final clip and never copy example timestamps:

```
{
  "has_demand": true,
  "segments": [
    {
      "start_ms": 11840,
      "end_ms": 15200,
      "transcript": "User's exact words",
      "structured_intent": "The user wants to...",
      "required_context": [
        {
          "type": "visual",
          "description": "User points to a bottle of olive oil"
        }
      ],
      "user_in_frame": false,
      "user_gender": "male",
      "user_age_group": "adult"
    }
  ]
}
If no demand: {"has_demand": false, "segments": []}
```

F.2 AUDIO-ONLY VARIANT

The audio-only instruction preserves the final-turn scope, temporal segmentation, transcript boundaries, self-contained intent, and JSON-only output requirements. It changes the input wording from video to audio and uses the following modality-specific fields and evidence rules.

Component	Audio-only specification
Demand evidence	Linguistic content and audio cues distinguish assistant-directed demands from other conversation.
Context types	Audio, dialogue history, user activity, user identity, user emotion, and scene condition; visual context is unavailable.
User profile fields	Output gender and age group; omit `user_in_frame`.

F.3 TEXT-ONLY INPUT STRUCTURE

The transcript baseline uses the shared prediction instruction above, with each media block replaced in place by ASR text. App. D.2 provides the full input templates, including dialogue-history framing and the no-speech replacement, together with complete positive and no-demand input examples and model settings. No reference transcript or demand label is supplied as input.

F.4 KEY-POINT JUDGE

Each reference point is supplied with its identifier, tier label, and point text.

Key-Point Judge

PURPOSE	Scores key-point coverage. For each key point the judge decides whether the predicted demand text covers it, under a lenient criterion that accepts paraphrase and hypernym coverage but still rejects a referent the prediction never names.
INPUTS	The predicted demand text for one segment, and that segment's key points.
OUTPUT	A JSON array with one entry per key point, carrying `qid`, `hit` and a short `reason`.

You are an evaluation expert. Given a passage of model output text and a set of key assessment points, judge for each assessment point whether the model output "hits" it.

Hit criterion (lenient):

- "hit" = true: the model output semantically covers the core of that assessment point, or that core can be reasonably inferred from / obtained as an equivalent restatement of the output – even if the wording differs, the phrasing is more general, or only the main point is covered while incidental modifiers are omitted.
- "hit" = false: only when the model output does not touch the core information of that assessment point at all, or substantively contradicts / gets it wrong.

Judgment principles:

- Each assessment point is independent; judge them one by one, looking only at whether that point's own core is covered.
- Hypernym / generalized hits are allowed: if the output covers the point with a more general statement (for example, "complaining about mold" covers "the mold is unusual this year"), it counts as a hit.
- When in doubt, be lenient: if the intent / context semantically covers the point on the whole, judge hit; judge miss only when the core content really is missing or is stated incorrectly.
- But do not invent what is absent: a distinctive referred-to object / action / attribute that is simply not in the output and cannot be reasonably inferred from it (for example, a specific "one particular bag" or "the region indicated by 'this one'" is never named) is still judged miss.

Output a JSON array in which each element has the following format:

```
{"qid": "kp01", "hit": true, "reason": "a brief justification of
    the hit or miss"}
```

Notes:

- You must give a judgment for every assessment point
- reason must be concise and state the key basis for the hit/miss
- Do not use double quotes inside reason; use single quotes or no quotes
- Output only the JSON array, with no other text